QUANTASPIKE: SHORT-WINDOW SPIKE-DRIVEN QUANTIZATION FOR LARGE LANGUAGE MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Large language models (LLMs) achieve strong performance across many tasks but rely on dense multiply-accumulate (MAC) operations during inference, resulting in high energy cost. Spiking neural networks (SNNs) offer an event-driven alternative in which synaptic integration uses lightweight accumulation. However, spike-driven LLM inference remains difficult because outlier-heavy activations typically require long firing windows or auxiliary non-spiking paths. We propose QuantaSpike, a short-window spike-driven quantization framework for LLMs built around Logarithmic Ternary Integrate-and-Fire (LTIF) neurons. LTIF uses ternary events with power-of-two membrane-response quanta, improving the information represented by each firing step while retaining shift-ACC-compatible computation. QuantaSpike combines this neuron with group-adaptive gain and selective outlier admission: normal values use residual LTIF steps, whereas admitted outliers receive one additional onset spike before entering the same residual dynamics. Across OPT and Llama-2, QuantaSpike achieves state-of-the-art or competitive perplexity and zero-shot accuracy among spike-driven LLM quantization methods. It also transfers to newer dense LLMs, remaining close to the FP16 reference on Llama-3-8B and Qwen3-8B under the same four-step firing window. Analytical linear-energy projections show that QuantaSpike reduces the energy of one linear transformation by about 80.0% on OPT models and 67.1% on Llama-2 models relative to SpikeQuant, providing an accurate and energy-efficient spike-driven path for LLM inference.

1 INTRODUCTION

In recent years, large language models (LLMs) have become a central foundation for modern Artificial Intelligence systems, driving advancements in natural language understanding, complex reasoning, code generation, and autonomous tool use (Brown et al., 2020; Ouyang et al., 2022; Chowdhery et al., 2022; Schick et al., 2023; Touvron et al., 2023a). Although these advances, LLM inference still relies on massive dense multiply-accumulate (MAC) operations, making energy-efficient deployment a persistent challenge. Spiking neural networks (SNNs), inspired by the brain's event-driven computation, provide an alternative: they carry information through sparse spike events and replace dense MAC operations with lightweight accumulate (ACC) operations, offering high biological plausibility and natural hardware efficiency (Horowitz, 2014; Wang et al., 2026a). This makes spike-driven LLM inference an appealing direction for reducing dense arithmetic in LLM deployment. However, representing LLM activations with spikes is difficult: LLM activations often exhibit heavy-tailed distributions and pronounced outliers that are critical to model quality (Dettmers et al., 2022; Xiao et al., 2023). Existing spike-driven LLM methods address this tension in two ways: extending the temporal firing window to improve activation fidelity (Xing

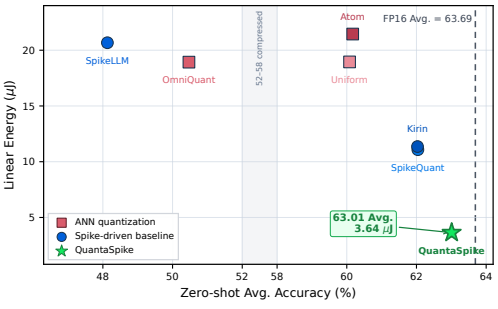


Figure 1: Zero-shot accuracy versus linear energy on Llama-2-7B.

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.

et al., 2025; Wang et al., 2026a), or introducing auxiliary integer-valued paths to handle activations that spikes cannot reliably represent (Wang et al., 2026b). However, the former increases inference latency and spike count, while the latter compromises the purity of spike-driven computation. The key challenge is therefore to maintain accuracy under a short firing window, while keeping the execution path spike-driven and shift-ACC compatible.

To address this challenge, we propose QuantaSpike, a short-window spike-driven quantization framework for LLMs. At its core is the Logarithmic LTIF neuron, which emits ternary spike values and uses power-of-two membrane-response quanta to represent values. QuantaSpike first determines onset admission using both statistical deviation and activation magnitude, then estimates group-adaptive gains from non-admitted values and initializes the membrane states according to the admission decisions. Normal values are represented by at most three residual LTIF firing steps, while admitted outliers obtain additional dynamic range through a single adaptive onset spike before entering the same residual LTIF dynamics. Within a firing window of at most four timesteps, the resulting events enable linear-layer computation through sign-controlled shifts and accumulation, without activation-side MAC operations. Across OPT, Llama-2, Llama-3, and Qwen3 models, QuantaSpike achieves competitive perplexity and zero-shot accuracy, while energy analysis shows that its short firing window and zero-spike silence reduce the number of ACC events actually executed.Our contributions are summarized as follows:

- We propose a short-window spike-driven quantization aiming to preserve both normal values and outliers without long firing windows or auxiliary non-spiking computation paths.
- We introduce QuantaSpike, a post-training framework centered on the LTIF neuron. LTIF combines ternary spike decisions with power-of-two membrane-response quanta within a four-step firing window.
- We achieve a strong accuracy–energy trade-off across OPT, Llama-2, Llama-3, and Qwen3 models. QuantaSpike delivers state-of-the-art spike-driven accuracy while substantially cutting energy via shift-ACC execution and zero-spike silence under the same analytical energy model used by prior work.

2 Related Work

LLM Quantization. Model quantization reduces the precision of weights and activations to lower the memory and computation costs of LLM inference. Post-training quantization (PTQ) has become a practical choice for LLMs, including weight-only methods such as GPTQ (Frantar et al., 2023) and AWQ (Lin et al., 2024), as well as weight–activation quantization methods such as SmoothQuant (Xiao et al., 2023), RPTQ (Yuan et al., 2023), and OmniQuant (Shao et al., 2024). Quantization-aware training further improves low-bit accuracy (Liu et al., 2024), while recent rotation-based methods such as QuaRot (Ashkboos et al., 2024) and SpinQuant (Liu et al., 2025) alleviate activation outliers for more aggressive quantization. Despite these advances, quantized LLMs still rely on dense low-bit activation representations and integer arithmetic.

Spike-Driven LLM Inference. Spike-driven LLMs aim to reduce the dense arithmetic cost of LLM inference by representing intermediate values with spike events and executing them through event-driven integration. Existing quantization-based approaches to spike-driven LLM inference broadly fall into two categories: temporal spike encoding and hybrid integer–spike representations. The former line of work follows temporal spike coding, where LLM activations are encoded into spike sequences and the resulting spiking dynamics are calibrated to approximate pretrained language models (Xing et al., 2025; Wang et al., 2026a; Chen et al., 2026; 2025). SpikeLLM uses generalized integrate-and-fire neurons and saliency-based spike allocation to reduce spike length, while SpikeQuant adopts TTFS-based activation encoding with salient-value handling. LAS introduces specialized neuron representations for outliers and nonlinear operators, while FAS uses coarse-to-fine calibration to reduce conversion error. These methods improve spike-based LLM inference, but their representation accuracy remains coupled to the firing window or to additional neuron-level encoding complexity.

The latter line of work relaxes the purely spike-driven representation by introducing hybrid precision for difficult activations (Wang et al., 2026b). Kirin observes that compact spike sequences struggle to preserve outlier activations and uses mixed-precision integer-spike representations to improve

reconstruction. This strategy reduces the burden on temporal spike coding, but weakens the uniform spike-driven execution path. QuantaSpike explores a different point: LTIF maintains accuracy under a short firing window, while keeping the runtime activation representation compatible with shift-ACC execution.

3 PRELIMINARIES

Integrate-and-fire (IF) neurons accumulate input current without a leak term and emit a spike when the membrane potential reaches a firing threshold. Their discrete-time dynamics are written as

$$\begin{cases} s^t = \mathbb{I}\big(H^{t-1} + I^t \geq \theta\big), \\ H^t = H^{t-1} + I^t - s^t\theta. \end{cases} \tag{1}$$

Here I^t is the input current, θ is the firing threshold, and H^{t-1} and H^t are the membrane states before and after the update. The first equation determines spike emission, while the second describes input integration and the membrane reset following a spike.

Some neurons also form autaptic connections, in which an axon terminal makes a synaptic contact with the same neuron's soma or proximal dendrites. For example, autaptic feedback is common in cortical GABAergic fast-spiking basket cells and has also been observed in a subset of layer-5 pyramidal cells (van der Loos & Glaser, 1972; Bacci et al., 2003). Neurotransmitter release at this self-synaptic contact generates an excitatory or inhibitory postsynaptic response and changes the membrane state of the same neuron after firing.

The spike response model (SRM) abstracts this temporal process by representing the membrane potential as the combination of an input-driven component and a response generated by the neuron's previous spikes. This formulation provides the basis for describing the spike-triggered membrane responses of LTIF (Gerstner, 1995; Gerstner et al., 2014).

4 METHOD

In this section, we introduce QuantaSpike, a spike-driven activation quantization workflow built around Logarithmic Ternary Integrate-and-Fire (LTIF) neurons. Figure 2 gives an overview of the proposed workflow. To avoid the long firing windows or mixed non-spiking paths used by prior spike-driven LLM quantization methods, QuantaSpike builds the activation representation around LTIF neurons. We first describe spike-aware activation mapping and the admission rule in Section 4.1, then formulate the LTIF membrane dynamics in Section 4.2. Section 4.3 shows how the resulting ternary spikes are executed through shift-ACC accumulation.

4.1 SPIKE-AWARE ACTIVATION MAPPING

As shown in Fig. 2, QuantaSpike maps LLM activations into initial membrane states for LTIF firing. The mapping combines two-criterion onset admission with group-wise scaling, allowing ordinary values and admitted outliers to enter the neuron with different dynamic-range requirements.

Given a token activation vector $x \in \mathbb{R}^H$, we partition the hidden dimension into contiguous groups of size G:

$$\mathcal{G}_m = \{j \mid mG \leq j < \min((m+1)G, H)\}. \tag{2}$$

Grouping preserves the original channel order and defines the scope of the local quantization scales.

Unlike prior spike-driven methods, we use a strict onset admission rule based jointly on statistical deviation and activation magnitude. For each token, we compute the median and median absolute deviation over the hidden dimension:

$$\mu = \mathrm{median}_j(x_j), \qquad D = \mathrm{median}_j\,|x_j - \mu|. \tag{3}$$

An activation becomes a candidate when its normalized deviation exceeds λ:

$$r_i = \mathbb{I}\left[\frac{|x_i - \mu|}{1.4826\max(D, \epsilon)} > \lambda\right]. \tag{4}$$

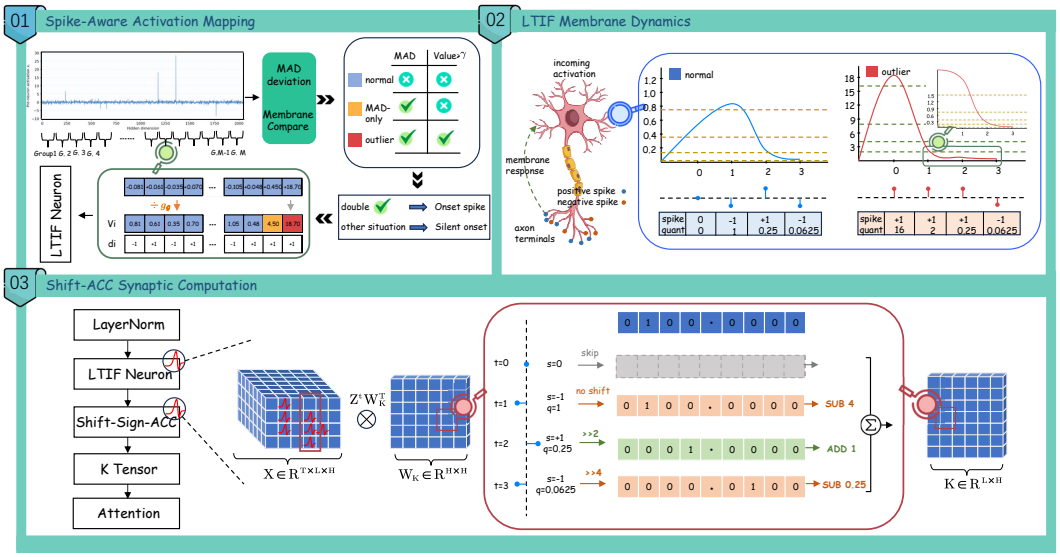


Figure 2: **Overview of QuantaSpike.** QuantaSpike maps LLM activations into the LTIF membrane domain, admits only necessary outlier states to an adaptive onset spike, and represents activations with short-window ternary LTIF spikes. These spikes are processed via sign selection, power-of-two shifts, and accumulation, ensuring a purely spike-driven linear path free of MAC operations.

We admit the candidate for onset firing only if its magnitude also exceeds γ:

$$a_i = r_i \, \mathbb{I}[|x_i| > \gamma]. \tag{5}$$

This second criterion prevents small but statistically unusual values from receiving an unnecessary onset event.

For each group, we estimate the normal-path gain from values with $a_j = 0$:

$$g_m = \max\left(\frac{\max_{j \in \mathcal{G}_m : a_j = 0} |x_j|}{Q_{\mathrm{n}}}, \epsilon \right), \qquad Q_{\mathrm{n}} = 2^{b_{\mathrm{n}}-1} - 1. \tag{6}$$

Excluding admitted outliers prevents extreme values from determining the scale of the normal path. The membrane is initialized as

$$h_i = \begin{cases} x_i / g_{\mathcal{G}(i)}, & a_i = 0, \\ \mathcal{Q}_{\mathrm{o},\mathcal{G}(i)}(x_i), & a_i = 1, \end{cases} \qquad u_i^0 = h_i, \tag{7}$$

where $\mathcal{Q}_{\mathrm{o},m}$ denotes group-wise outlier quantization followed by reconstruction in the original activation coordinates. Thus, the admission decision determines whether the initial state follows the normal path or the outlier path.

4.2 Logarithmic Ternary Integrate-and-Fire Neuron

Given the membrane initialization and admission decision from Section 4.1, LTIF represents each activation through signed ternary events and power-of-two membrane responses. The membrane state is initialized as $u_i^0 = h_i$, and its dynamics are

$$\begin{cases} s_i^t \in \{-1, 0, +1\}, \\ u_i^{t+1} = u_i^t - s_i^t q_i^t = h_i + \sum_{\tau=0}^{t} \eta_i(t+1, \tau) s_i^\tau. \end{cases} \tag{8}$$

where s_i^t is the ternary firing event and q_i^t is its non-negative response quantum. For the implemented LTIF response,

$$\eta_i(t+1, \tau) = -q_i^\tau, \qquad 0 \le \tau \le t. \tag{9}$$

Thus, each emitted event introduces a signed membrane correction that is retained in the residual membrane state. A positive event decreases a positive membrane state, whereas a negative event increases a negative membrane state toward zero. The membrane direction can reverse after an update when the residual crosses zero.The admission indicator directly gates the optional onset event: only values with $a_i = 1$ emit at $t = 0$, whereas all values enter the shared residual dynamics for $t \geq 1$.

LTIF neurons have three defining properties—signed ternary firing, graded thresholds, and admission-gated onset firing—which together enable them to represent the full activation range within a four-step firing window.

Signed ternary firing. LTIF emits a signed ternary spike

$$s_i^t \in \{-1, 0, +1\}, \tag{10}$$

where $+1$ and -1 denote positive and negative spike events, respectively, and 0 denotes a silent timestep. The sign records the direction of the current membrane state, while the effective spike magnitude is specified by an associated response factor. Each spike event is therefore paired with a non-negative power-of-two response magnitude:

$$q_i^t \in \{0\} \cup \{2^e : e \in \mathcal{E}\}, \tag{11}$$

where $\mathcal{E} \subset \mathbb{Z}$ is the supported exponent set. The ternary variable s_i^t determines the sign of the event, whereas q_i^t determines the strength of its membrane-response and its power-of-two contribution during synaptic computation.

Graded thresholds. LTIF uses a set of graded thresholds to control the response magnitude associated with each ternary spike. For group g and timestep t, calibration constructs an ordered threshold–response map

$$\mathcal{P}_g^t = \left\{ \left(\theta_{g,k}^t, q_{g,k}^t\right) \right\}_{k=1}^{K_t}, \qquad 0 < \theta_{g,1}^t < \cdots < \theta_{g,K_t}^t, \quad q_{g,k}^t = 2^{e_{g,k}^t}. \tag{12}$$

The thresholds determine the firing level reached by the membrane, while the corresponding response magnitude determine the strength of the signed membrane-response.

At time step t, following the LTIF membrane dynamics, the membrane magnitude $V_i^t = |u_i^t|$ is compared with the ordered thresholds, starting from the largest threshold. The highest threshold reached by the membrane determines the response level:

$$k_i^t = \max \left\{ k \,\middle|\, V_i^t \geq \theta_{\mathcal{G}(i),k}^t \right\}. \tag{13}$$

where $d_i^t = \operatorname{sgn}(u_i^t)$ (with $d_i^t = 0$ when $u_i^t = 0$), and the set is empty when $V_i^t < \theta_{\mathcal{G}(i),1}^t$. When V_i^t reaches at least the smallest threshold, LTIF emits a ternary spike in the current membrane direction and activates the corresponding signed membrane-response:

$$(s_i^t, q_i^t) = \begin{cases} \left(d_i^t, q_{\mathcal{G}(i),k_i^t}^t\right), & V_i^t \geq \theta_{\mathcal{G}(i),1}^t, \\ (0, 0), & V_i^t < \theta_{\mathcal{G}(i),1}^t. \end{cases} \tag{14}$$

Onset firing. According to the admission decision, LTIF actively controls the first timestep by either suppressing the onset event or emitting an adaptive onset spike. At $t = 0$, values with $a_i = 0$ are forced to remain silent. An admitted outlier receives one adaptive onset event:

$$s_i^0 = a_i \operatorname{sgn}(h_i), \qquad q_i^0 = a_i 2^{e_i^0}, \qquad e_i^0 = \arg\min_{e \in \mathcal{E}_\mathrm{o}} \left| \, |h_i| - 2^e \, \right|, \tag{15}$$

Here $\mathcal{E}_\mathrm{o}$ is the exponent set of the coarse outlier power-of-two bases after normalization by $g_{\mathcal{G}(i)}^\mathrm{o}(x)$; the bases are constructed from the observed outlier range. Thus, $q_i^0 = 0$ for non-admitted values. The onset event expands the representable range for admitted outliers, after which they enter exactly the same graded residual LTIF dynamics as normal values. The normalized spike-domain representation and its activation reconstruction are

$$z_i = \sum_{t=0}^{n} s_i^t q_i^t, \qquad \hat{x}_i = g_i(x) z_i, \qquad g_i(x) = \begin{cases} g_{\mathcal{G}(i)}^\mathrm{n}(x), & a_i = 0, \\ g_{\mathcal{G}(i)}^\mathrm{o}(x), & a_i = 1. \end{cases} \tag{16}$$

Normal values are represented using at most n residual firing steps, whereas admitted outliers use one additional onset step. With $n = 3$, QuantaSpike therefore uses at most four firing steps without introducing a separate non-spiking outlier path.

4.3 SHIFT-ACC COMPUTATION ANALYSIS

LTIF-coded activations can be executed through sign-controlled shift-and-ACC operations without activation-side MAC operations. Each nonzero event consists of a ternary sign and a power-of-two response quantum, while the group-adaptive gain is folded into the corresponding weight columns before inference.

For activation channel i, let z_i denote its normalized LTIF representation:

$$z_i = \sum_{t=0}^{n} s_i^t q_i^t, \qquad s_i^t \in \{-1, 0, +1\}, \quad q_i^t \in \{0\} \cup \{2^{e_i^t}\}. \tag{17}$$

The reconstructed activation is

$$\hat{x}_i = g_{\mathcal{G}(i)} z_i, \qquad \bar{W}_{:,i} = g_{\mathcal{G}(i)} W_{:,i}, \tag{18}$$

where $g_{\mathcal{G}(i)}$ is shared by the group containing channel i and is not the amplitude of an emitted spike. The gain is folded into the weight columns offline, so no activation-side multiplication is required at runtime.

For a linear layer $y = Wx$, substituting Eq.(18) gives

$$\hat{y} = \sum_i \bar{W}_{:,i} z_i = \sum_{i,t:s_i^t \neq 0} s_i^t \left(\bar{W}_{:,i} \ll e_i^t\right), \tag{19}$$

where $\ll e_i^t$ denotes multiplication by the power-of-two quantum $2^{e_i^t}$ through a bit shift. Thus, a positive event adds $\bar{W}_{:,i} \ll e_i^t$, a negative event subtracts the same shifted column, and a zero event performs no operation. Each nonzero LTIF event therefore requires one sign selection, one power-of-two shift, and one ACC update.

Zero-spike steps emit no event and trigger no accumulation. The runtime activation path consequently replaces activation-side MAC operations with sparse sign-controlled shift-ACC events, while preserving the reconstructed linear-layer output under the folded group gain.

5 EXPERIMENTS

5.1 EXPERIMENTAL SETUP

Implementation details. QuantaSpike employs a W4A4(4&5) post-training setting: 4-bit uniform weights and 4-bit activations with a 5-bit outlier budget. By default, we use a four-step firing window ($n = 3$ residual LTIF steps and one optional adaptive onset step) with a group size $G = 64$. We follow the standard 128-sequence, 2048-token WikiText-2 calibration protocol; the sampling and caching procedure is detailed in Appendix A. Experiments run on a single NVIDIA A100 80GB GPU.

Evaluation tasks. We evaluate QuantaSpike on OPT (Zhang et al., 2022) (1.3B, 2.7B), Llama-2 (Touvron et al., 2023b) (7B, 13B), Llama-3 (Grattafiori et al., 2024) (8B), and Qwen3 (Yang et al., 2025) (8B). Metrics include language modeling perplexity on WikiText-2 (Merity et al., 2017) and C4 (Raffel et al., 2020), and zero-shot accuracy on PIQA (Bisk et al., 2020), ARC-Easy, ARC-Challenge (Clark et al., 2018), HellaSwag (Zellers et al., 2019), and WinoGrande (Sakaguchi et al., 2020) following standard protocols.

Baselines. For OPT and Llama-2, we compare against ANN PTQ baselines (Vanilla, GPTQ (Frantar et al., 2023), OmniQuant (Shao et al., 2024), Atom (Zhao et al., 2024)) and spike-driven methods (SpikeLLM (Xing et al., 2025), SpikeQuant (Wang et al., 2026a), Kirin (Wang et al., 2026b)). For Llama-3 and Qwen3, where spike-driven results are limited, we evaluate FP16, uniform W4A4(4&5), and ANN baselines (GPTQ, AWQ (Lin et al., 2024)) alongside QuantaSpike under a unified pipeline with identical checkpoints, tokenizers, and evaluation setups.

5.2 MAIN RESULTS

In this section, we focus on whether QuantaSpike can maintain accuracy under a short firing window. Following prior spike-driven LLM studies, we report WikiText-2 and C4 perplexity for token-level language modeling, followed by zero-shot accuracy on five commonsense tasks. As shown in Table 1, QuantaSpike uses the shortest firing window among the compared spike-driven methods while maintaining competitive perplexity. On OPT models, it matches or improves upon SpikeQuant on both datasets and achieves the best WikiText-2 perplexity among quantized baselines on OPT-2.7B. On Llama-2 models, its perplexity remains close to SpikeQuant and Kirin despite reducing the firing window from 16 or 32 steps to four, indicating that the LTIF onset-and-residual representation preserves language-modeling quality.

Table 1: **Perplexity on WikiText-2(WT-2) and C4 across model scales.**We compare SpikeQuant (SQ), SpikeLLM (SL) and Kirin with our method QuantaSpike(QS)

Metric	OPT-1.3B				OPT-2.7B				Llama-2-7B					Llama-2-13B				
	FP16	SQ	Kirin	QS	FP16	SQ	Kirin	QS	FP16	SL	SQ	Kirin	QS	FP16	SL	SQ	Kirin	QS
Bits	W16 A16	W4A (4&5)	W4A (4&8)	W4A (4&5)	W16 A16	W4A (4&5)	W4A (4&8)	W4A (4&5)	W16 A16	W4A (4&8)	W4A (4&5)	W4A (4&8)	W4A (4&5)	W16 A16	W4A (4&8)	W4A (4&5)	W4A (4&8)	W4A (4&5)
Timestep	–	32	16	4	–	32	16	4	–	256	32	16	4	–	256	32	16	4
WT-2 ↓	14.63	16.38	**16.33**	16.38	12.47	13.15	12.89	**12.76**	5.68	11.36	5.80	**5.71**	5.78	4.88	9.71	5.04	**5.03**	5.11
C4 ↓	14.72	16.56	**16.26**	16.34	13.17	13.91	**13.72**	13.79	7.08	15.87	7.33	7.25	**7.23**	6.46	12.10	6.65	**6.61**	**6.61**

Table 2 further reports zero-shot accuracy on five commonsense tasks. QuantaSpike achieves the best average score on three of the four model settings, and is only 0.05 points behind Kirin on OPT-2.7B. Despite using a four-step firing window, QuantaSpike remains close to the FP16 reference, with only a 0.68-point average drop on both OPT-2.7B and Llama-2-7B. It also consistently outperforms the ANN quantization baselines in average accuracy, while preserving a spike-driven execution path. On Llama-2 models, QuantaSpike is particularly stable across PIQA, ARC-Easy, HellaSwag, and WinoGrande, showing that the shorter firing window does not translate into a systematic loss on downstream tasks. Together with the perplexity results, these findings support the central claim that QuantaSpike maintains accuracy under a short firing window while keeping activation computation spike-driven.

We further evaluate newer dense LLM families in Table 3. On Llama-3-8B, QuantaSpike obtains 6.98 WikiText-2 perplexity and a 71.25 zero-shot average, improving over GPTQ by 0.28 perplexity points and 1.11 zero-shot points. Its zero-shot score is also within 0.09 points of Uniform and 0.08 points of AWQ. On Qwen3-8B, QuantaSpike obtains 10.04 perplexity and a 69.41 zero-shot average; its perplexity is within 0.33 of FP16 and lower than those of AWQ and GPTQ. Compared with FP16, the zero-shot gaps are 1.48 points on Llama-3-8B and 2.17 points on Qwen3-8B. These results show that the LTIF representation transfers to newer dense LLMs under the same four-step firing window.

5.3 SHIFT-ACC ENERGY ANALYSIS

Each nonzero LTIF event is executed by sign selection, a power-of-two shift, and an ACC update, while silent residual steps trigger no accumulation. Let p_o denote the fraction of activations that emit an onset spike and p_z the fraction of silent residual steps. The average number of nonzero events per activation is

$$\bar{m} = n(1 - p_z) + p_o. \tag{20}$$

As detailed in Appendix C.1, accounting for silence reduces $\bar{m}$ from 3.03 to 1.34 on OPT-1.3B and from 3.01 to 2.08 on Llama-2-7B.

We estimate the energy of one linear transformation using the operation costs adopted by SpikeQuant and Kirin (Wang et al., 2026a;b). After normalizing all costs by a 4-bit ACC, the LTIF event cost and the resulting linear energy are

$$\begin{cases} C_{\mathrm{QS}} = n(1 - p_z)(1 + \alpha_{\mathrm{shift}}) + p_o(\rho_5 + \alpha_{\mathrm{shift}}), \\ E_{\mathrm{QS}}^{\mathrm{linear}}(M) = E_{\mathrm{SQ}}^{\mathrm{linear}}(M) \dfrac{C_{\mathrm{QS}}}{C_{\mathrm{SQ}}(M)}. \end{cases} \tag{21}$$

Table 2: **Zero-shot accuracy on OPT and Llama-2 models.** FP16 and QuantaSpike are evaluated in our pipeline; other baselines are reported from prior papers. Best and second-best results among quantized or spike-driven methods are bolded and underlined.

Model	Method	SNN Conv.	PIQA	ARC-Easy	ARC-Challenge	HellaSwag	WinoGrande	Avg.
OPT-1.3B	FP16	✗	72.52	50.84	29.78	53.72	59.75	53.32
	Vanilla Quant.	✗	67.79	42.93	25.26	45.94	54.93	47.37
	GPTQ	✗	68.34	46.17	27.65	46.53	55.26	48.79
	SpikeQuant	✓	71.49	**49.66**	27.22	51.46	**58.56**	51.68
	Kirin	✓	**71.86**	48.78	28.67	52.25	57.77	51.87
	QuantaSpike	✓	71.84	49.23	**28.79**	**52.34**	57.70	**51.98**
OPT-2.7B	FP16	✗	74.81	54.34	31.31	60.62	61.01	56.42
	Vanilla Quant.	✗	71.16	50.84	28.75	55.57	**60.46**	53.36
	GPTQ	✗	71.38	48.19	27.82	49.86	54.82	50.41
	SpikeQuant	✓	74.21	53.79	**31.66**	**59.11**	59.75	55.70
	Kirin	✓	**74.81**	**53.91**	31.48	58.83	59.91	**55.79**
	QuantaSpike	✓	74.55	53.84	30.97	59.05	60.30	55.74
Llama-2-7B	FP16	✗	78.51	55.54	42.33	73.99	68.06	63.69
	Vanilla Quant.	✗	75.46	53.96	40.19	69.21	61.56	60.08
	OmniQuant	✗	66.15	45.20	31.14	56.44	53.43	50.47
	Atom	✗	76.28	52.10	38.99	69.81	63.69	60.17
	SpikeLLM	✓	64.47	48.74	27.30	43.29	56.83	48.13
	SpikeQuant	✓	77.09	55.39	40.36	72.09	65.27	62.04
	Kirin	✓	76.77	54.55	40.53	72.10	66.22	62.03
	QuantaSpike	✓	**77.91**	**55.44**	**41.54**	**72.52**	**67.64**	**63.01**
Llama-2-13B	FP16	✗	80.52	57.53	44.98	77.38	70.22	66.13
	Vanilla Quant.	✗	76.71	55.13	41.47	73.09	65.19	62.32
	OmniQuant	✗	69.69	47.39	33.10	58.96	55.80	52.99
	Atom	✗	77.69	**57.58**	42.92	73.77	68.51	64.09
	SpikeLLM	✓	66.49	55.30	30.12	47.43	51.54	50.18
	SpikeQuant	✓	79.33	56.48	**44.03**	75.53	68.11	64.70
	Kirin	✓	79.35	56.40	43.94	75.94	68.98	64.92
	QuantaSpike	✓	**79.54**	56.63	43.18	**76.26**	**69.43**	**65.01**

Table 3: **WikiText-2 perplexity and zero-shot accuracy on newer dense LLMs.** Comparison of FP16, Uniform, AWQ, GPTQ and our QuantaSpike.

Method	Llama-3-8B					Qwen3-8B				
	FP16	Uniform	AWQ	GPTQ	QuantaSpike	FP16	Uniform	AWQ	GPTQ	QuantaSpike
SNN Conv.	✗	✗	✗	✗	✓	✗	✗	✗	✗	✓
WikiText-2 ↓	6.14	6.88	6.55	7.26	6.98	9.71	10.08	10.08	10.23	10.04
ZS Avg. ↑	72.73	71.34	71.33	70.14	71.25	71.58	70.01	70.66	70.59	69.41

where $\rho_5 = E^5_{\mathrm{ACC}}/E^4_{\mathrm{ACC}}$, $\alpha_{\mathrm{shift}} = E_{\mathrm{shift}}/E^4_{\mathrm{ACC}} = 0.2$, and $C_{\mathrm{SQ}}(M)$ is the corresponding normalized SpikeQuant cost.

Figure 3 shows that QuantaSpike has the lowest estimated linear-transformation energy across the evaluated OPT and Llama-2 models. Relative to SpikeQuant, it reduces linear energy by approximately 80.0% on OPT and 67.1% on Llama-2. These savings result from replacing dense activation-side MACs with sparse sign-controlled shift-ACC events and skipping accumulation during silent steps. Appendix C details the operation-cost assumptions and linear-energy projections.

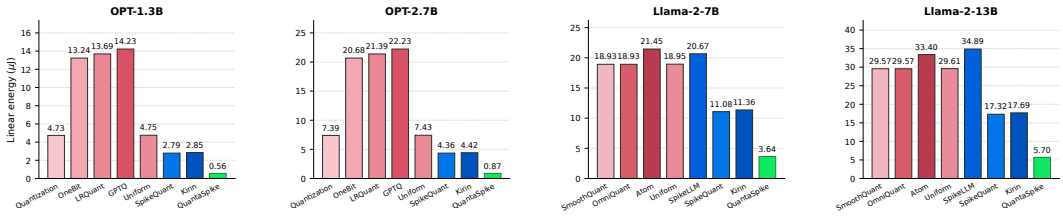


Figure 3: **Estimated energy of one linear transformation** under the operation-cost model adopted by SpikeQuant and Kirin.

5.4 Ablation Studies

We ablate two design choices in QuantaSpike: selective onset admission and the onset spike itself. Removing the magnitude gate leaves MAD-based admission alone, whereas removing the onset spike shortens the firing window from four to three timesteps and forces all values, including outliers, into the residual LTIF dynamics.

Table 4: **Ablation studies.** The onset spike improves outlier reconstruction, while the magnitude gate controls false outlier admission.

Ablation	Model	Variant	PPL	p_o	$\bar{m}$	Ratio
Onset spike	OPT-1.3B	QuantaSpike	16.38	2.90%	1.34	0.262
		w/o onset spike	18.78	0.00%	1.18	0.230
	Llama-2-7B	QuantaSpike	5.78	0.77%	2.08	0.407
		w/o onset spike	6.04	0.00%	1.95	0.382
Magnitude gate	Llama-2-7B	QuantaSpike	5.78	0.77%	2.08	0.407
		w/o magnitude gate	39.47	4.40%	2.27	0.445

Selective onset admission. As shown in Table 4, MAD-only admission increases the onset-spike rate on Llama-2-7B from 0.77% to 4.40%, while degrading WikiText-2 perplexity from 5.78 to 39.47. Statistical deviation alone therefore admits small-magnitude values that consume additional events without improving accuracy. The magnitude gate prevents these values from entering the onset-enabled representation.

Onset firing. Removing the onset spike saves one timestep but eliminates the dedicated response for admitted outliers. Consequently, WikiText-2 perplexity increases from 16.38 to 18.78 on OPT-1.3B and from 5.78 to 6.04 on Llama-2-7B, with a larger degradation on OPT-1.3B. Together, these results support selectively admitting outliers to an onset spike rather than broadly admitting MAD-selected values or eliminating onset firing.

6 Conclusion

In this paper, we presented QuantaSpike, a short-window spike-driven quantization framework for LLMs. QuantaSpike introduces the Logarithmic Ternary Integrate-and-Fire (LTIF) neuron, which combines signed ternary firing with power-of-two membrane-response quanta to increase the representation capacity of each firing step. Together with group-adaptive gain and selective onset admission, LTIF represents normal values and outliers through shared residual dynamics within a four-step window, while retaining an activation-side shift-ACC path without dense MAC operations. Across OPT, Llama-2, Llama-3, and Qwen3, QuantaSpike achieves state-of-the-art or competitive perplexity and zero-shot accuracy among spike-driven LLM quantization methods. Analytical energy estimates further show that QuantaSpike reduces linear-transformation energy by about 80.0% on OPT and 67.1% on Llama-2 relative to SpikeQuant, supporting accurate and energy-efficient spike-driven LLM inference. The limitations and future directions are discussed in Appendix F.

7 Reproducibility Statement

Appendix A specifies the quantization settings, calibration data and sampling seed, evaluation protocol, and model-specific hyperparameters. The remaining appendices provide full task results, ablations, firing-event accounting, and the assumptions behind the analytical energy estimates. QuantaSpike is a post-training method and does not update the pretrained model weights. Our reported perplexity and zero-shot scores are single evaluations of each quantized checkpoint under a fixed calibration seed. FP16, uniform quantization, and QuantaSpike results generated in our pipeline are distinguished from baseline results taken from published reports, which retain the source papers' evaluation settings. We will release the source code, configurations, calibration and evaluation scripts, and plotting scripts upon publication.

8 Ethics Statement

Our experiments use publicly available pretrained language models and benchmark datasets; this work collects no new human data. We use these resources for research under the terms and documentation provided by their original publishers. Quantization can make LLM inference less costly, but it does not remove the potential for biased, inaccurate, or harmful model outputs. Applications using quantized models should therefore retain appropriate evaluation and safeguards for their intended setting. The efficiency results in this paper are operation-level estimates, consistent with the practices of prior works, and should be interpreted within the stated accounting scope.

References

Saleh Ashkboos, Amirkeivan Mohtashami, Maximilian L. Croci, Bo Li, Pashmina Cameron, Martin Jaggi, Dan Alistarh, Torsten Hoefler, and James Hensman. Quarot: Outlier-free 4-bit inference in rotated llms. In *Advances in Neural Information Processing Systems*, volume 37, 2024.

Alberto Bacci, John R. Huguenard, and David A. Prince. Functional autaptic neurotransmission in fast-spiking interneurons: A novel form of feedback inhibition in the neocortex. *Journal of Neuroscience*, 23(3):859–866, 2003. doi: 10.1523/JNEUROSCI.23-03-00859.2003.

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

Tom B. Brown et al. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, 2020.

Long Chen, Xiaotian Song, Andy Song, BaDong Chen, Jiancheng Lv, and Yanan Sun. FAS: Fast ANN-SNN conversion for spiking large language models. *arXiv preprint arXiv:2502.04405*, 2025.

Long Chen, Xiaotian Song, and Yanan Sun. LAS: Loss-less ANN-SNN conversion for fully spike-driven large language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, 2026.

Aakanksha Chowdhery et al. PaLM: Scaling language modeling with pathways. *arXiv preprint arXiv:2204.02311*, 2022.

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. Think you have solved question answering? try ARC, the AI2 reasoning challenge. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018.

Tim Dettmers, Mike Lewis, Younes Belkada, and Luke Zettlemoyer. LLM.int8(): 8-bit matrix multiplication for transformers at scale. In *Advances in Neural Information Processing Systems*, volume 35, 2022.

Elias Frantar, Saleh Ashkboos, Torsten Hoefler, and Dan Alistarh. GPTQ: Accurate post-training quantization for generative pre-trained transformers. In *International Conference on Learning Representations*, 2023.

Wulfram Gerstner. Time structure of the activity in neural network models. *Physical Review E*, 51 (1):738–758, 1995. doi: 10.1103/PhysRevE.51.738.

Wulfram Gerstner, Werner M. Kistler, Richard Naud, and Liam Paninski. *Neuronal Dynamics: From Single Neurons to Networks and Models of Cognition*. Cambridge University Press, 2014.

Aaron Grattafiori et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

Mark Horowitz. 1.1 computing's energy problem (and what we can do about it). In *2014 IEEE International Solid-State Circuits Conference Digest of Technical Papers (ISSCC)*, pp. 10–14. IEEE, 2014. doi: 10.1109/ISSCC.2014.6757323.

Ji Lin, Jiaming Tang, Haotian Tang, Shang Yang, Wei-Ming Chen, Wei-Chen Wang, Guangxuan Xiao, Xingyu Dang, Chuang Gan, and Song Han. AWQ: Activation-aware weight quantization for on-device llm compression and acceleration. In *Proceedings of Machine Learning and Systems*, volume 6, pp. 87–100, 2024.

Zechun Liu, Barlas Oguz, Changsheng Zhao, Ernie Chang, Pierre Stock, Yashar Mehdad, Yangyang Shi, Raghuraman Krishnamoorthi, and Vikas Chandra. Llm-qat: Data-free quantization aware training for large language models. In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 467–484, 2024.

Zechun Liu, Changsheng Zhao, Igor Fedorov, Bilge Soran, Dhruv Choudhary, Raghuraman Krishnamoorthi, Vikas Chandra, Yuandong Tian, and Tijmen Blankevoort. Spinquant: Llm quantization with learned rotations. In *International Conference on Learning Representations*, 2025.

Stephen Merity, Caiming Xiong, James Bradbury, and Richard Socher. Pointer sentinel mixture models. In *International Conference on Learning Representations*, 2017.

Long Ouyang et al. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems*, volume 35, 2022.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67, 2020.

Keisuke Sakaguchi, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. WinoGrande: An adversarial winograd schema challenge at scale. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pp. 8732–8740, 2020. doi: 10.1609/aaai.v34i05.6399.

Timo Schick, Jane Dwivedi-Yu, Roberto Dessi, Roberta Raileanu, Maria Lomeli, Eric Hambro, Luke Zettlemoyer, Nicola Cancedda, and Thomas Scialom. Toolformer: Language models can teach themselves to use tools. In *Advances in Neural Information Processing Systems*, volume 36, pp. 68539–68551, 2023. doi: 10.52202/075280-2997.

Wenqi Shao, Mengzhao Chen, Zhaoyang Zhang, Peng Xu, Lirui Zhao, Zhiqian Li, Kaipeng Zhang, Peng Gao, Yu Qiao, and Ping Luo. OmniQuant: Omnidirectionally calibrated quantization for large language models. In *International Conference on Learning Representations*, 2024.

Hugo Touvron et al. LLaMA: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023a.

Hugo Touvron et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023b.

H. van der Loos and E. M. Glaser. Autapses in neocortex cerebri: Synapses between a pyramidal cell's axon and its own dendrites. *Brain Research*, 48:355–360, 1972. doi: 10.1016/0006-8993(72)90189-8.

Chenyu Wang, Zhanglu Yan, Zhi Zhou, Xu Chen, and Weng-Fai Wong. Energy-efficient and dequantization-free quantization of LLMs: A spiking neural network approach to salient value mitigation. In *Proceedings of the ACM Web Conference 2026*, pp. 5334–5345, 2026a. doi: 10.1145/3774904.3792337.

Chenyu Wang, Zhanglu Yan, Zhi Zhou, Xu Chen, and Weng-Fai Wong. Kirin: Improving ANN efficiency with SNN hybridization. *arXiv preprint arXiv:2602.08817*, 2026b.

Guangxuan Xiao, Ji Lin, Mickael Seznec, Hao Wu, Julien Demouth, and Song Han. SmoothQuant: Accurate and efficient post-training quantization for large language models. In *Proceedings of the 40th International Conference on Machine Learning*, 2023.

Xingrun Xing, Boyan Gao, Zheng Liu, David A. 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.

An Yang et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.

Zhihang Yuan, Lin Niu, Jiawei Liu, Wenyu Liu, Xinggang Wang, Yuzhang Shang, Guangyu Sun, Qiang Wu, Jiaxiang Wu, and Bingzhe Wu. Rptq: Reorder-based post-training quantization for large language models. *arXiv preprint arXiv:2304.01089*, 2023.

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

Susan Zhang et al. OPT: Open pre-trained transformer language models. *arXiv preprint arXiv:2205.01068*, 2022.

Yilong Zhao, Chien-Yu Lin, Kan Zhu, Zihao Ye, Lequn Chen, Size Zheng, Luis Ceze, Arvind Krishnamurthy, Tianqi Chen, and Baris Kasikci. Atom: Low-bit quantization for efficient and accurate LLM serving. In *Proceedings of Machine Learning and Systems*, volume 6, 2024.

A Additional Experimental Details

A.1 Quantization and Calibration

QuantaSpike is applied after pretraining and does not update model weights. All quantized models use the W4A4(4&5) setting described in the main text: weights are quantized to 4-bit uniform values, normal activation states use a 4-bit LTIF representation, and admitted outliers use one additional onset firing step, corresponding to a 5-bit representation budget. The residual LTIF path uses $n = 3$ firing steps unless otherwise specified, and the optional onset response increases the maximum firing window to four steps.

For an activation vector, QuantaSpike partitions the hidden dimension into contiguous groups of size G and computes token-dependent normal and outlier gains for each group. To match the common protocol used by prior data-driven PTQ and spike-driven LLM methods, calibration uses 128 sequences of 2048 tokens randomly sampled from the WikiText-2 training split with seed 0. The calibration sequences are disjoint from the evaluation split and are tokenized with the corresponding model tokenizer. Hooks capture the inputs to every target linear layer over these sequences, and the captured tensors are concatenated for calibration. The calibration activations determine the normal LTIF prototypes and exponent range. At evaluation time, the median and median absolute deviation (MAD) are recomputed for each token to identify outlier candidates, and the magnitude gate determines whether a candidate receives the optional onset spike.

Table 5: **Model-specific QuantaSpike configurations.** n is the number of residual LTIF firing steps, G is the group size, λ is the MAD coefficient, and γ is the magnitude-gate coefficient.

Model family	n	G	λ	γ	Max. window
OPT	3	64	3.0	0.5	4
Llama-2	3	64	3.0	0.5	4
Llama-3	3	32	3.0	0.5	4
Qwen3 dense	3	32	3.0	0.6	4
Qwen3 MoE	3	32	3.0	0.6	4

Table 7: **Full results on newer dense LLMs.** We report WikiText-2 perplexity and zero-shot accuracy. Normalized accuracy is used for PIQA, ARC-Easy, ARC-Challenge, and HellaSwag; accuracy is used for WinoGrande. All three methods are evaluated in the same pipeline.

Model	Method	PPL ↓	PIQA	ARC-Easy	ARC-Challenge	HellaSwag	WinoGrande	Avg. ↑
Llama-3-8B	FP16	6.14	80.85	77.69	53.33	79.15	72.61	72.73
	Uniform	6.88	79.49	77.15	50.43	77.79	71.82	71.34
	QuantaSpike	6.98	80.15	77.05	50.11	77.47	71.47	71.25
Qwen3-8B	FP16	9.71	77.75	80.93	56.57	74.93	67.72	71.58
	Uniform	10.08	75.90	76.14	53.41	73.72	65.90	69.01
	QuantaSpike	10.04	76.42	76.55	54.72	73.19	66.18	69.41

All calibration-dependent quantities, including group gains, MAD statistics, and LTIF reset prototypes, are fixed before evaluation. QuantaSpike therefore requires neither task-specific fine-tuning nor gradient updates. We cache the calibrated prototypes using the model, layer, number of residual firing steps, group size, bit width, and exponent range as the cache key. Changing any of these quantities invalidates the corresponding cache entry.

Table 5 lists the configurations used for the reported results. OPT and Llama-2 use the default setting. For Llama-3-8B and Qwen3-8B, a smaller group size better matches their local activation statistics while preserving the same four-step firing window.

A.2 Evaluation Protocol

Perplexity is evaluated on the WikiText-2 test split and the English C4 validation split, using each model's tokenizer and non-overlapping 2048-token windows. For C4, we follow the data-construction convention in prior spike-driven LLM evaluation: tokenize the concatenation of the first 1,100 documents from the first English validation shard (`en/c4-validation.00000-of-00008.json.gz`), retain the first 256×2048 tokens, and compute perplexity over the resulting 256 windows. This C4 evaluation data is separate from the WikiText-2 training sequences used for calibration. FP16, uniform quantization, and QuantaSpike are evaluated on the same windows within our pipeline; values copied from prior papers remain under their published protocols. Zero-shot evaluation is performed with the same version of the LM Evaluation Harness across all methods evaluated in our pipeline. We report normalized accuracy for PIQA, ARC-Easy, ARC-Challenge, and HellaSwag, and accuracy for WinoGrande. The reported average is the unweighted mean over these five tasks. All post-training quantization and evaluation experiments are conducted on a single NVIDIA A100 80GB GPU.

B Full Results on Newer Dense LLMs

Prior spike-driven LLM studies primarily report results on OPT and Llama-2. We therefore evaluate Llama-3-8B and Qwen3-8B against FP16 and uniform low-bit quantization in the same pipeline. Table 7 gives the complete zero-shot results, extending the averages reported in the main text. QuantaSpike remains close to uniform quantization on both model families under a four-step firing window. The selected group size is particularly important for Llama-3-8B, where the final QuantaSpike configuration achieves 6.98 WikiText-2 perplexity and a 71.25 zero-shot average. On Qwen3-8B, setting the admission gate to $\gamma = 0.6$ yields a 69.41 zero-shot average while retaining $n = 3$ residual steps.

Table 6: **Hyperparameter sensitivity on newer dense LLMs.** The selected four-step configurations are marked in bold.

Model	n	G	γ	WikiText-2 PPL ↓
Llama-3-8B	3	64	0.5	7.211
	3	**32**	**0.5**	**6.979**
	3	32	0.6	7.002
	4	32	0.5	6.958
Qwen3-8B	3	64	0.5	10.109
	3	32	0.5	10.046
	3	**32**	**0.6**	**10.037**
	3	32	0.7	10.055
	4	32	0.5	10.046

B.1 Sensitivity to Group Size, Admission Gate, and Firing Window

Table 6 reports the compact search used for the newer dense models. Reducing G from 64 to 32 consistently improves perplexity, indicating that finer gain alignment is more important than increasing the residual firing budget for these models. Increasing n from 3 to 4 at fixed $G = 32$ and $\gamma = 0.5$ gives only a small additional reduction, from 6.979 to 6.958 on Llama-3-8B and from 10.048 to 10.046 on Qwen3-8B. We therefore retain $n = 3$ in the main configuration to preserve the four-step firing window. For Qwen3-8B, $\gamma = 0.6$ provides the best balance among the tested admission thresholds; increasing it to 0.7 slightly worsens perplexity.

C Energy Accounting Details

C.1 Firing Budget

We quantify the firing budget by counting nonzero LTIF spikes rather than the maximum number of available firing steps. Let p_o denote the fraction of activation values that emit an adaptive onset spike, and let p_z denote the fraction of silent steps among all residual LTIF steps. For a window containing n residual steps and one optional onset step, the average number of nonzero events per activation is

$$\bar{m} = n(1 - p_z) + p_o. \tag{22}$$

The first term counts nonzero residual events, while the second accounts for onset events emitted by admitted outliers. We use $n = 3$ for the results reported here. Each nonzero event triggers an ACC update for each associated synaptic connection, whereas a silent step triggers no accumulation.

To isolate the accumulation cost associated with this event count, we report the normalized activation-side arithmetic ratio

$$R_{\text{act}} = \frac{\bar{m} E_{\text{ACC}}}{E_{\text{MAC}}}, \tag{23}$$

where $E_{\text{ACC}} = 0.9\,\text{pJ}$ and $E_{\text{MAC}} = 4.6\,\text{pJ}$. The reference is one MAC per activation–weight interaction. This simplified ratio accounts for accumulation only; it excludes shift operations, memory access, and other execution overheads. It is therefore an event-based arithmetic proxy, rather than a measurement of end-to-end hardware energy. The bit-dependent ACC costs and shift costs used for the linear-energy projection are treated separately in the following subsection.

To quantify the contribution of zero-spike silence, we compare the actual nonzero-event count with an accounting reference that treats every residual step as active. This reference retains the same onset rate, LTIF representation, and reconstructed values, but sets $p_z = 0$ for accounting purposes, giving

$$\bar{m}_{\text{all}} = n + p_o. \tag{24}$$

It does not modify the quantizer or constitute a separate inference experiment; perplexity is unchanged by construction. The reduction in counted events is

$$\bar{m}_{\text{all}} - \bar{m} = n p_z. \tag{25}$$

Table 8 reports the firing statistics used in the energy analysis. On OPT-1.3B, 56.37% of residual steps are silent, reducing the average counted events per activation from 3.03 to 1.34. On Llama-2-7B, 30.92% of residual steps are silent, reducing the count from 3.01 to 2.08. The corresponding normalized arithmetic ratios decrease from 0.593 to 0.262 and from 0.588 to 0.407, respectively. These results show how data-dependent silence reduces the actual event count within the four-step firing window.

C.2 Linear-Energy Projection

The absolute linear-energy comparison in Fig. 3 uses the operation-level accounting adopted by the prior linear-layer analyses cited in Section 5.3. This projection concerns one linear transformation and should be interpreted as an analytical estimate rather than a measurement of end-to-end hardware power.

Table 8: **Effect of zero-spike silence on event accounting.** All residual steps counted is an accounting reference that treats silent residual steps as active; it preserves the same representation and perplexity as QuantaSpike. p_o is the onset-event rate, p_z is the silent residual-step rate, $\bar{m}$ is the average counted events per activation, and R_{act} is the normalized accumulation-cost ratio.

Model	Accounting	p_o	p_z	$\bar{m}$	R_{act}
OPT-1.3B	All residual steps counted	2.90%	0.00%	3.03	0.593
	QuantaSpike (nonzero only)	2.90%	56.37%	**1.34**	**0.262**
Llama-2-7B	All residual steps counted	0.77%	0.00%	3.01	0.588
	QuantaSpike (nonzero only)	0.77%	30.92%	**2.08**	**0.407**

For QuantaSpike, each nonzero residual LTIF event consists of one power-of-two shift and one 4-bit ACC update. An admitted outlier additionally emits one onset event, whose response is accumulated with the corresponding 5-bit ACC cost. We normalize all operation costs by the energy of a 4-bit ACC and write the normalized QuantaSpike event cost as

$$C_{\mathrm{QS}}(M) = n(1 - p_z)(1 + \alpha_{\mathrm{shift}}) + p_o(\rho_5 + \alpha_{\mathrm{shift}}), \tag{26}$$

where n is the number of residual firing steps, p_z is the fraction of silent residual steps, and p_o is the fraction of values that emit an adaptive onset spike. The quantity

$$\rho_5 = \frac{E^5_{\mathrm{ACC}}}{E^4_{\mathrm{ACC}}}$$

denotes the relative cost of a 5-bit ACC, and

$$\alpha_{\mathrm{shift}} = \frac{E_{\mathrm{shift}}}{E^4_{\mathrm{ACC}}} = 0.2$$

denotes the normalized cost of a power-of-two shift. The residual term therefore accounts for nonzero 4-bit LTIF spikes, whereas the onset term accounts for the additional 5-bit event emitted by admitted outliers.

For the SpikeQuant reference, we use the same operation-energy accounting as the source linear-layer analysis. Let ρ_M denote the reported salient-value ratio for model M, and let μ_{n} and μ_{o} denote the normalized MAC costs for the normal and salient-value paths, respectively. The corresponding SpikeQuant cost is

$$C_{\mathrm{SQ}}(M) = (1 - \rho_M)(1 + \mu_{\mathrm{n}}) + \rho_M(\rho_5 + \mu_{\mathrm{o}}), \tag{27}$$

where the numerical values used in the projection are $\mu_{\mathrm{n}} = 8.66$ and $\mu_{\mathrm{o}} = 9.24$. These values follow the normalized operation costs in the reference analysis, while ρ_M follows the corresponding reported salient-value ratio.

We project the QuantaSpike energy by scaling the reported SpikeQuant linear-transformation energy according to the ratio of the two normalized operation costs:

$$E^{\mathrm{linear}}_{\mathrm{QS}}(M) = E^{\mathrm{linear}}_{\mathrm{SQ}}(M)\frac{C_{\mathrm{QS}}(M)}{C_{\mathrm{SQ}}(M)}. \tag{28}$$

Here $E^{\mathrm{linear}}_{\mathrm{SQ}}(M)$ is the reported SpikeQuant energy for one linear transformation, and $E^{\mathrm{linear}}_{\mathrm{QS}}(M)$ is the corresponding QuantaSpike projection under the same accounting model.

For OPT-1.3B and Llama-2-7B, the values of p_o and p_z are obtained from the measured firing statistics in Table 8. For OPT-2.7B and Llama-2-13B, we reuse the measured statistics from the smaller model in the same model family because exact firing diagnostics were not recorded for those models. The resulting QuantaSpike values are therefore marked as estimates in Table 9.

Table 9 lists the numerical values used to generate Fig. 3. The projected QuantaSpike linear energy is approximately 80.0% lower than SpikeQuant on OPT models and 67.1% lower on Llama-2 models under this operation-level accounting. These percentages describe the estimated energy of a single linear transformation. They do not represent measured end-to-end energy, and they do not include attention matrix multiplications, memory-system overheads, or hardware-specific scheduling effects.

Table 9: **Linear-energy values used in Fig. 3** (μJ). Prior-method values follow the corresponding reported linear-layer analysis. QuantaSpike values are obtained from the operation-level projection above. * indicates an estimate using same-family firing statistics. Lower is better.

Method	OPT-1.3B	OPT-2.7B	Llama-2-7B	Llama-2-13B
Quantization	4.732	7.393	–	–
OneBit	13.237	20.683	–	–
LRQuant	13.690	21.390	–	–
GPTQ	14.227	22.229	–	–
SmoothQuant	–	–	18.927	29.573
OmniQuant	–	–	18.927	29.573
Atom	–	–	21.454	33.398
SpikeLLM	–	–	20.668	34.887
Uniform (ours)	4.753	7.426	18.948	29.606
SpikeQuant	2.790	4.358	11.082	17.318
Kirin	2.851	4.424	11.363	17.688
QuantaSpike	**0.558**	**0.872***	**3.643**	**5.695***

D Additional Ablation Results

D.1 Adaptive Onset Spike

The onset-spike ablation tests whether outliers can be represented by the residual LTIF dynamics alone. Removing the onset response yields only a modest reduction in the event ratio, from 0.262 to 0.230 on OPT-1.3B and from 0.407 to 0.382 on Llama-2-7B. The reconstruction penalty is substantially larger: outlier MSE increases by $16.3\times$ on OPT-1.3B and $6.3\times$ on Llama-2-7B. The corresponding perplexity degradation shows that the onset event is not redundant firing overhead. It provides the additional membrane correction required to return admitted outliers to the operating range of the shared residual LTIF dynamics.

Table 10: **Ablation of the adaptive onset spike.** Removing the onset response saves a small event budget but sharply increases outlier reconstruction error.

Model	Variant	PPL	Outlier MSE	p_o	$\bar{m}$	Ratio
OPT-1.3B	QuantaSpike	**16.380**	**0.00326**	2.90%	1.337	0.262
	w/o onset	18.784	0.05322	0.00%	**1.178**	**0.230**
Llama-2-7B	QuantaSpike	**5.779**	**0.00334**	0.77%	2.082	0.407
	w/o onset	6.039	0.02093	0.00%	**1.950**	**0.382**

D.2 Magnitude Gate

MAD alone identifies values that are unusual relative to their channel distribution, but not every such value lies outside the gain-aligned LTIF range. As shown in Table 11, removing the magnitude gate increases the onset rate from 0.77% to 4.40% on Llama-2-7B, yet perplexity rises sharply from 5.78 to 39.47. This behavior is therefore not an energy–accuracy trade-off. Without the magnitude gate, false admissions alter the initial membrane correction for values that are already representable by the residual LTIF prototypes, thereby disrupting the calibrated reconstruction path. The magnitude gate thus serves as an admission mechanism rather than a conventional sparsity heuristic.

Table 11: **Ablation of the magnitude gate on Llama-2-7B.** MAD-only admission spends more onset events but severely degrades reconstruction.

Variant	PPL	p_o	$\bar{m}$	Ratio
QuantaSpike	**5.779**	**0.77%**	**2.082**	**0.407**
w/o magnitude gate	39.470	4.40%	2.270	0.445

E Diagnostic Extension to Mixture-of-Experts Models

We evaluate whether the LTIF representation can be applied to routed expert activations using Qwen3-30B-A3B. Unlike a dense model, an MoE model exposes each expert to only the tokens

selected by the router. Expert-wise calibration is therefore uneven: frequently selected experts receive many calibration values, whereas rarely selected experts may receive too few valid samples to estimate stable gains and reset prototypes.

We consider two scopes. The attention setting applies QuantaSpike to attention and shared dense projections while keeping routed experts and the router in high precision. The expert setting additionally applies QuantaSpike to routed expert projections, while the router remains in high precision. To test whether calibration coverage is the limiting factor, we also use a more diverse calibration set that activates a larger number of expert instances.

Table 12 reports a small-scale WikiText-2 diagnostic with eight sequences of length 512. Quantizing attention and shared dense projections preserves the FP16 perplexity on this split. Extending QuantaSpike to routed experts increases perplexity from 11.97 to 12.25. The diverse calibration set substantially reduces the number of skipped expert calibration entries but does not change perplexity at this evaluation scale. These results establish functional compatibility with routed expert computation, but they should not be interpreted as a full WikiText-2 benchmark.

Table 12: **Diagnostic MoE results on Qwen3-30B-A3B.** The router remains in high precision. Evaluation uses eight WikiText-2 sequences of length 512; lower perplexity is better.

Variant	Quantized scope	Calibrated	Skipped	PPL
FP16	None	–	–	11.970
Attention	Attn. + shared dense	192	0	**11.900**
Experts	Routed experts	10,578	7,854	12.254
Experts + diverse calib.	Routed experts	13,488	4,944	12.254

The skipped entries arise from incomplete routing coverage during calibration rather than from the LTIF quantizer itself. A larger calibration corpus or routing-aware sample selection would improve expert coverage, but a full MoE study is outside the scope of the dense-model evaluation in the main paper.

F Limitations and Future Work

Limitations. Our main evaluation focuses on dense decoder-only LLMs, including OPT, Llama-2, Llama-3, and Qwen3. The MoE experiment in Appendix E is a small diagnostic rather than a full routed-expert benchmark: expert activation counts are uneven, several expert-specific calibration entries remain incomplete, and the evaluation uses only a short WikiText-2 split. We do not evaluate vision-language models, where visual-token statistics, modality-specific outliers, and cross-modal projection layers may require different admission and calibration rules.

Future work. Future work will develop routing-aware calibration for MoE models, with sample allocation that provides sufficient coverage for rarely selected experts and separate statistics for shared and expert projections. We also plan to extend LTIF to vision-language models by studying modality-aware gains and onset admission for visual and textual activations.

G The Use of Large Language Models

Large language models were used to assist with literature retrieval, text editing, and the preparation of plotting scripts during the writing of this paper. The authors reviewed, verified, and edited all generated content and take full responsibility for the final manuscript, including all reported claims, data, and citations.