CONTEXT-ROBUST: CONTEXT-INFORMED ACTIVATION CALIBRATION FOR MIXTURE-OF-EXPERTS LANGUAGE MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Sparse routing makes activation calibration in Mixture-of-Experts (MoE) models inherently router-conditioned: each expert observes only the calibration tokens routed to it, so a small activation range observed in one calibration realization does not necessarily indicate a stable small range. Consequently, conventional Absmax calibration can select overly fine quantization steps that are sensitive to the routed calibration population. We introduce **Context-Robust (CR)**, a training-free activation calibration method that uses contextualized natural-routing executions to probe routed range instability. CR measures channelwise routed-range discrepancies between original and contextualized calibration collections, aggregates them into a site-level routed range-instability signal, and uses this signal as a one-sided floor on channelwise Absmax quantization steps. The resulting steps remain fixed during inference. Across Qwen3.5-35B-A3B and Gemma-4-26B-A4B, CR consistently improves average downstream performance over Absmax across FP16, AWQ, GPTQ, and SmoothQuant weight configurations, while exhibiting a repeatable task-selective preservation pattern. Controlled W16A8 diagnostics further show substantially reduced sample-level degradation dispersion and upper-tail failures while keeping mean degradation close to Absmax. These results show that router-conditioned range instability provides useful calibration information beyond activation magnitude for MoE activation quantization.

1 INTRODUCTION

Post-training quantization (PTQ) is widely used to reduce the memory and inference cost of large language models. For Mixture-of-Experts (MoE) models, activation quantization inside routed expert feed-forward networks (FFNs) is especially relevant because the selected experts still account for substantial inference computation even though only a small subset is active per token.

Most activation quantization methods, such as SmoothQuant (Xiao et al., 2023), derive fixed quantization steps from statistics collected on a calibration set. In a dense FFN, all calibration tokens contribute to the same layer statistics. In an MoE, routing filters those observations: for calibration set D and expert set $S(x)$ selected for token x, expert e only observes

$$D_e = \{x \in D \mid e \in S(x)\}.$$

Both the size and composition of D_e vary substantially across experts. Consequently, an expert's activation range is determined jointly by the calibration content and the model's routing decisions. Simply forcing additional samples through an expert is not equivalent, because it changes the naturally routed distribution the expert is intended to process (Chen et al., 2025b). We therefore view MoE activation calibration as a *router-conditioned calibration problem*.

This distinction is critical for range-based calibration in MoE. In conventional Absmax quantization, the step size is proportional to the largest activation observed for a channel, so a small observed maximum yields a fine step. Such a fine step is appropriate only if the routed activation range is stable. However, a small maximum from a single calibration run could not establish stability. A different run under natural routing can alter hidden path, expert assignments, and hence the routed population from which the range is estimated. The key issue is not only the magnitude of the

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.

observed activations but also the sensitivity of the routed range estimate to the calibration run. An unstable range estimate can lead to overly fine steps for some channels and clipping resolution in deployment. Therefore, popular MoE PTQ approaches like GPTQ and PTQ (Frantar et al., 2023; Lin et al., 2024) only perform weight-only quantization (W4A16), i.e. keeping the activations at full precision. To address the uneven expert routing distribution across calibration datasets inherent in conventional methods, MoEQuant (Chen et al., 2025b) introduces Expert-Balanced Self-Sampling. However, self-sampling incurs significant computational overhead and yields no performance gains when using comparable amount of tokens.

Beyond measuring activation magnitude, we seek to probe how internal activations of a token can vary while the model's task behavior remains largely unchanged. Common text augmentations include word substitution, deletion, paraphrasing, and back-translation. Although intended to preserve meaning and model output, they change some target tokens. Here we keep the target text fixed and vary its preceding context: for each target A_i, we prepend natural-language contexts B_j to form B_jA_i. Experiments in Appendix A show that such prefixes can substantially change internal representations while preserving the model's task decision for most evaluated pairs. Motivated by this observation, we compare activation ranges across original and contextualized executions under natural routing to inform calibration.

Based on this view, we introduce **Context-Robust (CR)**, a training-free activation calibration method for post-dispatch expert FFNs. At each expert quantization site, CR compares channelwise routed activation ranges with and without prefixes and takes the median of their absolute discrepancies to obtain the site-level routed range-instability signal τ_k. CR then uses τ_k as a lower bound on the quantization step:

$$\Delta_{\mathrm{CR},k,c} = \max(\Delta_{\mathrm{abs},k,c}, \tau_k)\,.$$

Here, $k = (l, e, s)$ indexes a post-dispatch activation quantization site. The indices l and e identify the MoE layer and expert, respectively. The site type s denotes either the shared input to the gate/up projections or the input to the down projection, and c indexes an activation channel at that site. The conventional Absmax step $\Delta_{\mathrm{abs},k,c}$ is computed from the maximum absolute activation observed for this channel in the Absmax calibration collection, i.e. symmetric quantization. This maximum is divided by the largest positive quantization integer at the chosen bit width. The resulting $\Delta_{\mathrm{CR},k,c}$ is the fixed CR quantization step for that channel. CR leaves an Absmax step unchanged if it is at least τ_k; otherwise, it raises the step to τ_k. This retains coverage of the observed activation range and prevents channels from using very small steps when the observed ranges change substantially after adding prefixes. Contextualized executions are used only during calibration, and inference uses ordinary fixed steps.

Our experiments support three conclusions. First, across Qwen3.5-35B-A3B under FP16, AWQ, GPTQ, and SmoothQuant weight configurations, CR consistently improves average downstream performance over Absmax, while exhibiting a repeatable but non-uniform taskwise preservation pattern. Second, experiments on Gemma-4-26B-A4B show that the same calibration principle transfers to a different sparse-expert architecture. Third, reference-state and sample-level analyses show that CR favors selective preservation rather than uniformly reducing quantization error; matched-step and same-structure non-contextual controls further show that this behavior is not reproduced by generic step enlargement or by a shared-floor construction alone.

Our contributions are:

- We formulate MoE expert activation calibration as a router-conditioned range-estimation problem in which fine Absmax resolution can depend strongly on the routed calibration realization.
- We introduce CR, which compares activation ranges collected with and without contextual prefixes under natural routing, then uses the discrepancies to set a shared minimum quantization step within each expert quantization site.
- We validate CR across multiple weight configurations and two sparse-expert architectures, and provide sample-level and token-level analyses that characterize its selective-preservation behavior and distinguish it from generic step inflation, shared-floor structure alone, and overly strong interpretations of the channelwise routed-range discrepancy.

2 Related Work

Activation calibration and sparse-expert quantization. PTQ methods use calibration data to set quantization parameters through range estimation, clipping, reconstruction, or transformations of weights and activations. SmoothQuant rescales weights and activations to reduce the difficulty of quantizing activation outliers. GPTQ and AWQ focus on weight quantization, using reconstruction error and activation information, respectively. In MoE models, each expert receives only the calibration tokens selected by the router. MoEQuant uses expert-balanced sampling and affinity guidance (Chen et al., 2025b). EAQuant combines expert-aware smoothing, routing-consistency alignment, and calibration balancing (Fu et al., 2026). EAC-MoE and mixed-precision methods also consider expert selection or sensitivity (Chen et al., 2025a; Chowdhury et al., 2026). CR addresses a different aspect of calibration: whether the activation ranges observed under natural routing remain stable when the calibration context changes. It uses these range changes to set activation quantization steps, without changing the routing policy or assigning different bit widths to different experts.

Prediction consistency under input perturbations. Prior work studies whether model predictions remain stable under perturbations intended to preserve task labels. CheckList formalizes this through invariance tests, including the addition of irrelevant context (Ribeiro et al., 2020). For sparse MoE language models, MoE-RBench evaluates adversarial and out-of-distribution robustness, including word-level transformations and sentence-level style shifts, and reports advantages over dense counterparts under the evaluated fine-tuning settings (Chen et al., 2024). Irrelevant information can distract hidden states of large language models when it does not alter the correct answer (Shi et al., 2023). Our work connects this behavioral perspective to activation calibration rather than robustness training. We empirically examine decision preservation under contextual prefixes (Appendix A), then use original and contextualized executions to probe router-conditioned activation-range variation. CR uses these range discrepancies to construct fixed quantization steps, without imposing an output-consistency training objective.

3 Problem Formulation

Quantization scope. Let h_l be the representation entering MoE layer l, $E_{l,e}(h_l)$ the output of expert e, $S_l(h_l)$ the top-K experts selected by the router, and $\alpha_{l,e}(h_l)$ the normalized routing weight assigned to expert e. The routed branch is

$$y_l = \sum_{e \in S_l(h_l)} \alpha_{l,e}(h_l) E_{l,e}(h_l).$$

We quantize only post-dispatch activations inside naturally selected experts: the shared input to the gate/up projections and the gated activation entering the down projection. The router, router input, attention modules, and other non-target operations remain in high precision. We index a quantization site by $k = (l, e, s)$ with $s \in \{\text{gate/up}, \text{down}\}$. Because quantization is applied after dispatch, it cannot alter the expert set at the same layer, although its perturbation can propagate to later layers.

Router-conditioned range estimation. Let $\mathcal{X}_k(\mathcal{D})$ be the activations that naturally reach site k when calibration collection $\mathcal{D}$ is executed under the original routing policy. Channelwise Absmax calibration uses

$$M_{k,c}(\mathcal{D}) = \max_{x \in \mathcal{X}_k(\mathcal{D})} |x_c|, \qquad \Delta^{(n)}_{\text{abs},k,c}(\mathcal{D}) = \frac{M_{k,c}(\mathcal{D})}{q^{(n)}_{\max}}.$$

A small observed maximum yields a fine quantization grid, but also leaves little representational headroom beyond the values seen during calibration. For signed static n-bit quantization, the largest positive value representable by channel c is $q^{(n)}_{\max}\Delta_{k,c}$. Under Absmax,

$$q^{(n)}_{\max}\Delta^{(n)}_{\text{abs},k,c} = M_{k,c},$$

so the representable range is tied directly to the maximum observed in that particular calibration realization. If this maximum is unusually small because only a limited routed population was observed, Absmax simultaneously chooses a very fine grid and provides no additional headroom for larger activations reached under another natural-routing realization; such values are then more likely

to clip. Thus, fine resolution is desirable when the routed range estimate is stable, but can be overly dependent on the calibration realization when that estimate is unstable. In an MoE, $\mathcal{X}_k(\mathcal{D})$ depends on both calibration content and routing, so a small $M_{k,c}$ does not by itself imply a reliably small routed range.

Consider original and contextualized calibration collections

$$\mathcal{D}_A = \{A_i\}, \qquad \mathcal{D}_{BA} = \{B_j A_i\},$$

both executed under natural routing. In general,

$$\mathcal{X}_k(\mathcal{D}_A) \neq \mathcal{X}_k(\mathcal{D}_{BA}),$$

because adding prefixes can change both the activations and which tokens reach each expert. We therefore compare the activation ranges collected with and without prefixes while keeping the routing policy unchanged. The following geometric view provides an intuition for why such context-induced variation can be informative for calibration.

Geometric intuition. Consider a fixed representation location in the model and a target token occurrence v. Let $\mathcal{C}_v$ denote the set of representation vectors of this target occurrence obtained under the different prefix contexts considered in calibration. For two representation vectors, we use $\mathrm{dist}(\cdot,\cdot)$ to denote their distance. The within-context diameter of v is

$$\omega_v = \sup_{z,z' \in \mathcal{C}_v} \mathrm{dist}(z, z').$$

Let $\mathcal{V}$ denote the collection of target token occurrences under consideration. We define the nearest cross-cluster separation of v as

$$\mu_v = \inf_{w \in \mathcal{V},\, w \neq v} \; \inf_{\substack{z \in \mathcal{C}_v \\ z' \in \mathcal{C}_w}} \mathrm{dist}(z, z').$$

When $\omega_v < \mu_v$, contextual variation and cross-cluster variation occur at distinct geometric scales. In this idealized separable regime, there exists a non-empty interval

$$\omega_v \leq \rho_v < \mu_v,$$

where ρ_v denotes an abstract representation-resolution scale, rather than the scalar quantization step Δ used by the actual quantizer. Intuitively, a resolution much finer than the within-context variation attempts to preserve distinctions that the model representation itself does not maintain across contexts, whereas a resolution approaching the nearest cross-cluster separation may begin to erase distinguishable representation structure.

In practice, calibration observes only finite subsets $\widehat{\mathcal{C}}_v \subseteq \mathcal{C}_v$ for $v \in \mathcal{V}$. We define the sampled within-cluster diameter

$$\widehat{\omega}_v = \max_{z,z' \in \widehat{\mathcal{C}}_v} \mathrm{dist}(z, z'),$$

and the sampled nearest cross-cluster separation

$$\widehat{\mu}_v = \min_{w \in \mathcal{V},\, w \neq v} \; \min_{\substack{z \in \widehat{\mathcal{C}}_v \\ z' \in \widehat{\mathcal{C}}_w}} \mathrm{dist}(z, z').$$

Since $\widehat{\mathcal{C}}_v \subseteq \mathcal{C}_v$, finite sampling gives

$$\widehat{\omega}_v \leq \omega_v, \qquad \mu_v \leq \widehat{\mu}_v.$$

Hence, in the separable regime,

$$\widehat{\omega}_v \leq \omega_v < \mu_v \leq \widehat{\mu}_v.$$

Finite calibration can therefore underestimate contextual variation while overestimating the separation to the nearest representation cluster. This makes the sampled inter-cluster separation an optimistic basis for choosing a maximal resolution scale, while the observed contextual variation provides a more conservative lower-scale signal against excessively fine quantization.

This geometric picture motivates, rather than directly defines, CR. CR does not explicitly construct token-level clusters or estimate ω_v and μ_v. Instead, it uses the variation of naturally routed activation ranges between original and contextualized executions as a streamable site-level proxy for calibration instability. Appendix A provides a qualitative visualization of the corresponding context-induced representation structure.

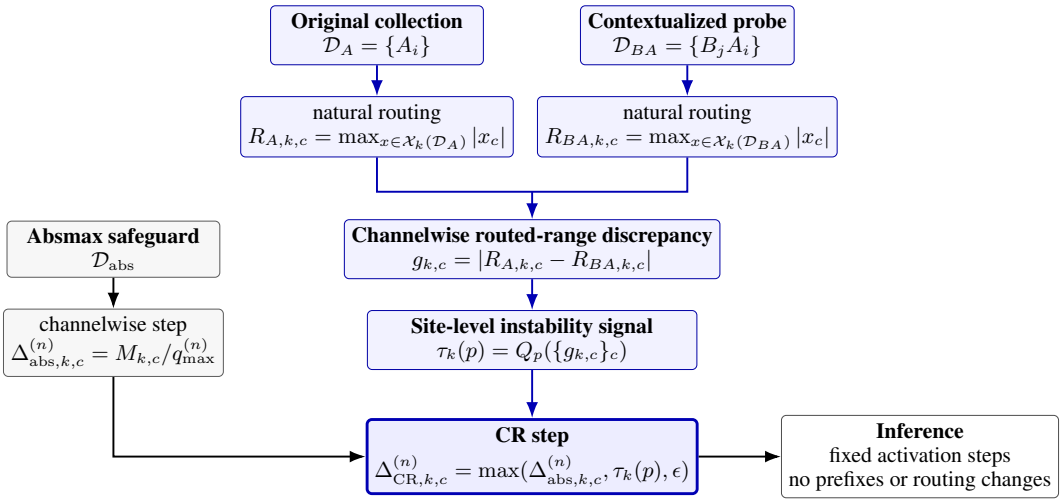

Figure 1: Overview of Context-Robust activation calibration. Original and contextualized inputs are executed under the unchanged natural-routing policy. Their collection-level routed ranges define channelwise discrepancies $g_{k,c}$, aggregated into a site-level routed range-instability signal τ_k that acts as a one-sided floor on the conventional channelwise Absmax step. The same token need not reach the same expert across executions, and the contextualized path is used only during calibration.

Design objective. The goal is not to estimate a token-level contextual displacement or simply to enlarge the representable range. Instead, we regularize overly fine channelwise resolution when the corresponding expert quantization site exhibits unstable routed range statistics. The calibration rule must be one-sided,

$$\Delta^{(n)}_{k,c} = F\left(\Delta^{(n)}_{\text{abs},k,c}, \tau_k\right), \qquad F(a,t) \geq a, \quad F(a,0) = a,$$

so context information cannot shrink a channel below its Absmax-derived step. Section 4 defines τ_k as a streamable site-level routed range-instability signal and realizes F as a shared floor.

4 Context-Robust Activation Calibration

Figure 1 shows the calibration pipeline. The original and contextualized paths probe how naturally routed range statistics change across alternative executions, while the Absmax path provides the conventional channelwise observed-range safeguard. The three paths are separated just by role. $\mathcal{D}_{\text{abs}}$ may coincide with $\mathcal{D}_A$. Contextualized inputs are used only during calibration, and all final steps are fixed for ordinary inference.

4.1 Contextualized Natural-Routing Probe

For each target A_i, CR forms contextualized executions $B_j A_i$. Original and contextualized collections are both run with the model's unchanged routing policy; CR does not force the same token through the same expert. At every observed site k and channel c, it accumulates the absolute routed ranges

$$R_{D,k,c} = \max_{x \in \mathcal{X}_k(\mathcal{D}_D)} |x_c|, \qquad D \in \{A, BA\}.$$

All contextualized executions are pooled before range accumulation, making the statistic streamable and avoiding storage of token-level activations. The notation distinguishes three calibration roles: $\mathcal{D}_A$ supplies the unprefixed routed ranges, $\mathcal{D}_{BA}$ supplies the contextualized probe ranges, and $\mathcal{D}_{\text{abs}}$ supplies the conventional Absmax safeguard. The method does not require these roles to use disjoint samples; in particular, $\mathcal{D}_{\text{abs}}$ can be identical to $\mathcal{D}_A$. Weight calibration and activation-calibration artifacts must be specified separately. Appendix H describes the calibration inputs and implementation details.

4.2 SITE-LEVEL ROUTED RANGE INSTABILITY

CR first measures the channelwise routed-range discrepancy,

$$g_{k,c} = |R_{A,k,c} - R_{BA,k,c}|,$$

and aggregates it across channels using

$$\tau_k(p) = Q_p(\{g_{k,c}\}_c)\,.$$

Unless stated otherwise, the context aggregation quantile is $p = 0.5$. Thus τ_k is one scalar per expert quantization site. It is an *empirical site-level routed range-instability signal*: because the ranges are computed over naturally routed finite populations, it may reflect contextual changes in hidden trajectories, routed-population changes, observation counts, and sampling variation of a max statistic. We do not interpret τ_k as token-level contextual displacement. Appendix E further characterizes the contributions of contextualized execution and finite routed-population effects.

The median summarizes the typical channel discrepancy at a site without allowing a few extreme channels to determine the shared floor. The independent Absmax term still protects each channel's observed range, while τ_k expresses instability shared by the site.

4.3 CONTEXT-INFORMED QUANTIZATION-STEP FLOOR

CR collects the conventional Absmax safeguard $M_{k,c}$ from $\mathcal{D}_{\text{abs}}$ and defines

$$\Delta^{(n)}_{\text{CR},k,c} = \max\left(\frac{M_{k,c}}{q^{(n)}_{\max}}, \tau_k(p), \epsilon\right).$$

The Absmax term retains coverage of the observed activation range, while τ_k sets a shared lower bound on the channelwise quantization steps within each expert quantization site. Note that ϵ stands for a small positive constant to avoid zero division. Channels whose Absmax steps meet the floor are unchanged; smaller steps are raised to the floor. At the default $p = 0.5$, the CR floor determines 98.865% of final step elements with Qwen3.5-35B-A3B; Absmax remains the channelwise coverage safeguard and fallback when no site-level signal is available. Thus, the observed activation magnitudes determine the Absmax steps, and the range differences measured with and without prefixes determine the shared minimum step.

By default, we use τ_k itself as the minimum quantization step, without a tuned multiplier. This choice does not imply that the measured range difference equals quantization error. More generally, we can multiply this lower bound by a floor-strength coefficient γ:

$$\Delta^{(\gamma)}_{k,c} = \max(\Delta_{\text{abs},k,c}, \gamma\tau_k)\,,$$

where γ controls the strength of the CR floor. We use $\gamma = 1$ in the main method to avoid introducing a tuned coefficient. The sensitivity study in Appendix B shows that both $\gamma = 0.5$ and $\gamma = 1$ reduce variance and P99 relative to Absmax, whereas $\gamma = 2$ substantially increases mean degradation and P99. Thus, the benefit is not limited to $\gamma = 1$, although these results do not establish it as an optimal setting.

Separately, the context aggregation quantile p controls how τ_k itself is estimated. Appendix D studies the effect of p, while Appendices B and C test whether CR can be reproduced by generic step enlargement or by the shared-floor structure alone.

CR is entirely post-training. All statistics can be accumulated as running maxima, and the final steps are fixed after calibration. Inference therefore requires no contextual prefixes, paired execution, routing modification, or online scale search. Appendix H gives the exact nominal construction and implementation caveats.

5 EXPERIMENTS

We first evaluate whether CR preserves downstream capability across weight configurations on Qwen3.5-35B-A3B, then test whether the same calibration principle transfers to Gemma-4-26B-A4B. We next analyze the selective-preservation behavior observed in the main results. Additional sample-level degradation analyses, alternative-explanation controls, calibration-signal diagnostics, and sensitivity studies are reported in the appendices.

5.1 Experimental Setup

Our primary model is Qwen3.5-35B-A3B, and our secondary model is Gemma-4-26B-A4B. We evaluate FP16 weights and W4 configurations based on AWQ, GPTQ, and a SmoothQuant-based configuration (SQ). FP16 weights with INT8 expert activations are denoted W16A8; the W4 configurations use W4A8 at the targeted sparse-expert operations. AWQ and GPTQ reconstruct W4 weights using the first 256 SkyPile examples for weight calibration in the CR and Absmax configurations, while SQ applies its smoothing transform from activation ranges. CR uses $p = 0.5$ unless stated otherwise.

The activation-calibration collections have different roles from the 256-example weight-calibration set. CR obtains its unprefixed ranges from 32 SkyPile target examples and its contextualized ranges from pairing those targets with 32 GSM8K training-example prefixes, yielding $32 \times 32 = 1{,}024$ naturally routed prefixed executions. An independent collection of 1,024 SkyPile examples supplies the channelwise Absmax safeguard for both CR and the Absmax baseline. The unprefixed targets are included in that larger SkyPile collection; the contextualized executions provide the additional site-level signal, not additional weight-calibration samples. Appendix H specifies token observation and site coverage.

Downstream tasks are MMLU (Hendrycks et al., 2021), BoolQ (Clark et al., 2019), HellaSwag (Zellers et al., 2019), OpenBookQA (Mihaylov et al., 2018), MathQA (Amini et al., 2019), GSM8K (Cobbe et al., 2021), and HumanEval (Chen et al., 2021). Appendices H and I provide the calibration, implementation, and evaluation details; the remaining appendices give the protocols for the supporting diagnostics and controls.

5.2 Main Results on Qwen3.5-35B-A3B

Table 1 reports the Qwen3.5-35B-A3B results. CR has the highest mean among activation-quantized methods in all four weight configurations, exceeding Absmax by 4.55, 6.14, 5.44, and 5.06 percentage points for AWQ, GPTQ, SQ, and FP16 weights, respectively. These unweighted seven-task averages recur across weight configurations rather than tracking one W4 reconstruction method. The gains are not uniform across tasks: MMLU, HellaSwag, and HumanEval improve in every configuration, BoolQ, MathQA, and GSM8K decline, and OpenBookQA is mixed.

HumanEval shows particularly large gains across all four weight configurations; Appendix G analyzes its prediction drift and clipping behavior, while Appendix F reports a calibration-collection resampling check on selected failures.

5.3 Cross-Architecture Validation

We next evaluate the Gemma-4-26B-A4B base checkpoint on the same seven tasks with the same Absmax/EBSS/CR comparison to test whether the calibration principle transfers across sparse-expert architectures. Gemma uses the evaluation budgets reported in Table 2, while the Qwen3.5 results in Table 1 use the full evaluation splits.

Table 2 shows that CR also has the highest activation-quantized mean on Gemma-4-26B-A4B in all four weight configurations. Relative to Absmax, the seven-task mean rises by 6.29, 6.88, 6.57, and 7.24 points for AWQ, GPTQ, SQ, and FP16 weights. The repeated ordering across architectures supports transfer of the calibration principle, although the largest gains are concentrated on GSM8K and HumanEval and the multiple-choice outcomes remain mixed.

5.4 Selective Preservation Analysis

CR is not a better global approximation of the reference output distribution. Across 585,698 token positions, Absmax has lower global mean $\Delta\mathrm{NLL}$ (0.029561 vs. 0.034006) and fewer vocabulary top-1 disagreements (7.66% vs. 10.54%). Figure 2 shows that the effect is strongly reference-state dependent: CR reduces degradation in high-confidence and large-margin regions, but perturbs uncertain states more strongly. For example, vocabulary top-1 disagreement drops from 0.47% to 0.24% in the highest-confidence bin and from 0.38% to 0.14% in the largest-margin bin, while increasing from 37.44% to 46.20% at the smallest margins.

Table 1: Qwen3.5-35B-A3B downstream task performance (%). AWQ, GPTQ, SQ, and FP16 identify weight configurations; Absmax, EBSS, and CR identify activation-calibration methods. EBSS is the external expert-balanced self-sampling baseline from MoEQuant (Chen et al., 2025b). All seven tasks use their complete evaluation splits. HS denotes HellaSwag and OBQA denotes OpenBookQA. HumanEval reports pass@1; the remaining tasks report accuracy.

Weight	Activation	W/A	MMLU	BoolQ	HS	OBQA	MathQA	GSM8K	HumanEval	Average
AWQ	Absmax	4/8	79.93	**89.57**	76.98	**30.80**	**54.81**	**93.48**	23.17	64.11
	EBSS	4/8	75.49	88.17	73.78	30.20	50.28	93.33	18.90	61.45
	CR **(Ours)**	4/8	**83.33**	87.37	**80.74**	29.80	53.40	92.95	**53.05**	**68.66**
GPTQ	Absmax	4/8	79.42	**89.14**	76.86	30.80	**54.20**	**94.16**	20.73	63.62
	EBSS	4/8	75.18	86.61	73.82	28.80	49.82	92.87	17.07	60.60
	CR **(Ours)**	4/8	**82.97**	86.85	**80.62**	**32.80**	53.07	93.48	**58.54**	**69.76**
SQ	Absmax	4/8	79.92	**89.33**	76.92	30.80	**54.71**	**94.16**	21.95	63.97
	EBSS	4/8	75.71	88.90	73.88	29.00	49.48	92.80	20.73	61.50
	CR **(Ours)**	4/8	**83.61**	87.61	**80.51**	**32.60**	53.94	93.33	**54.27**	**69.41**
FP16	Absmax	16/8	79.93	**89.48**	77.13	**31.40**	**55.41**	**93.86**	21.34	64.08
	EBSS	16/8	76.15	88.17	74.08	28.40	50.15	93.78	18.90	61.38
	CR **(Ours)**	16/8	**83.60**	86.73	**80.87**	30.80	54.34	93.40	**54.27**	**69.14**
Baseline	FP16	16/16	84.91	87.22	81.49	34.20	61.57	93.10	64.63	72.45

Table 2: Gemma-4-26B-A4B downstream task performance (%). Due to compute limits, tasks with sample counts below their full evaluation splits use the subset sizes shown in the last row.

Weight	Activation	W/A	MMLU	BoolQ	HS	OBQA	MathQA	GSM8K	HumanEval	Average
AWQ	Absmax	4/8	73.24	**83.40**	67.77	**35.80**	49.22	13.67	8.54	47.38
	EBSS	4/8	71.58	79.98	64.84	32.60	44.24	14.06	6.71	44.86
	CR **(Ours)**	4/8	**75.00**	82.52	**69.14**	35.20	**49.90**	**31.64**	**32.32**	**53.67**
GPTQ	Absmax	4/8	73.54	**83.01**	67.38	35.40	**50.00**	10.94	8.54	46.97
	EBSS	4/8	73.14	81.45	64.84	32.20	43.46	13.28	6.10	44.92
	CR **(Ours)**	4/8	**75.78**	81.25	**68.46**	**36.20**	49.02	**36.33**	**29.88**	**53.85**
SQ	Absmax	4/8	73.93	**83.30**	68.55	34.60	**50.29**	17.58	7.93	48.03
	EBSS	4/8	74.41	81.25	63.67	31.60	44.14	16.02	6.71	45.40
	CR **(Ours)**	4/8	**75.68**	83.01	**70.12**	**35.40**	48.54	**37.11**	**32.32**	**54.60**
FP16	Absmax	16/8	73.93	**83.50**	68.16	35.80	**51.95**	17.19	10.37	48.70
	EBSS	16/8	74.61	81.15	65.04	31.40	45.12	14.45	7.32	45.58
	CR **(Ours)**	16/8	**77.44**	82.23	**69.04**	**36.20**	49.80	**43.36**	**33.54**	**55.94**
Baseline	FP16	16/16	77.05	84.67	69.24	36.40	53.12	55.86	37.20	59.08
Evaluation samples per task			1024	1024	1024	500	1024	256	164	—

The pooled mean ΔNLL weights reference token positions, while downstream scores depend on scored decisions and generated continuations. Thus, lower pooled error can coexist with weaker task scores: CR redistributes error, reducing disagreement in confident states though it is higher at small margins.

The HumanEval W16A8 diagnostic provides a task-specific example (Appendix G). On 9,877 reference continuation tokens, CR lowers teacher-forced top-1 disagreement from 8.36% to 2.92% and mean ΔNLL from 0.18715 to 0.01929 relative to Absmax; mean first teacher-forced divergence shifts from 6.3 to 15.5 tokens. Despite higher gate/up input relative RMSE (0.13092 versus 0.05304), down-input clipping falls from 0.23876% to 0.01480%. These diagnostics are consistent with fewer clipping events and later prediction drift benefiting autoregressive generation, without requiring uniformly lower local reconstruction error.

Supporting degradation and control diagnostics. A controlled Qwen3.5-35B-A3B W16A8 analysis shows that CR keeps mean sample-level perplexity degradation close to Absmax while substantially reducing degradation dispersion and the upper tail. Matched step-inflation controls further show that generic enlargement of the Absmax steps does not reproduce CR; full sample-level statistics, matched-step controls, and floor-strength sensitivity are reported in Appendix B. A same-structure non-contextual control likewise shows that a shared-floor rule with matched total

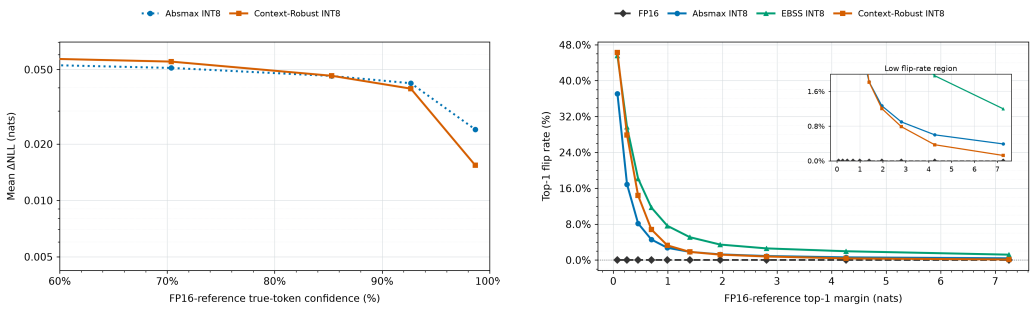

Figure 2: Reference-state-dependent behavior under INT8 expert activation quantization. Left: mean ΔNLL stratified by reference true-token confidence. Right: vocabulary top-1 disagreement rate stratified by reference log-probability margin. CR is not uniformly closer to the reference model; its relative advantage appears primarily in high-confidence and large-margin states, while uncertain states can be perturbed more strongly.

step mass is insufficient without the contextual site-wise allocation; the full protocol and results are in Appendix C.

Calibration-signal diagnostics and sensitivity. Routed-count-matched analyses indicate that τ_k contains information beyond an independent matched-count sampling null, while remaining sensitive to finite routed-population effects. We therefore interpret τ_k as a context-informed empirical site-level routed range-instability signal rather than a direct estimate of token-level contextual displacement; Appendix E gives the full analysis. Sensitivity studies further show that moderate floor strengths reduce degradation dispersion, whereas overly aggressive floors degrade both average and tail behavior. We use $p = 0.5$ as the median-based default operating point; the full floor-strength and context-aggregation-quantile studies are provided in Appendices B.1 and D.

6 Limitations and Future Work

The current τ_k statistic is an empirical routed-range-instability signal and can reflect both context-associated variation and finite-sample effects from routed populations. Separating these factors more precisely may lead to better calibration statistics. The current mapping from τ_k to the quantization-step floor is also deliberately simple, leaving adaptive or theoretically motivated mappings for future work. Finally, broader validation across MoE architectures and real low-precision kernels is needed to characterize generality and translate the calibration benefits into end-to-end efficiency gains.

7 Conclusion

We introduced Context-Robust (CR), a training-free activation calibration method for sparse Mixture-of-Experts models. CR treats expert activation calibration as a router-conditioned range-estimation problem and uses contextualized natural-routing executions to identify unstable routed activation ranges. Their channelwise range discrepancies are aggregated into a site-level signal and used as a one-sided floor on conventional Absmax quantization steps.

Across Qwen3.5-35B-A3B and Gemma-4-26B-A4B, CR improves average downstream performance over Absmax across multiple weight configurations. Our analyses further show that CR exhibits selective preservation rather than uniformly reducing approximation error: its relative advantage is concentrated in high-confidence and large-margin reference states, while matched-step and same-structure non-contextual controls show that the observed behavior is not reproduced by generic step enlargement or by the shared-floor structure alone. These results suggest that router-conditioned range instability provides useful calibration information beyond activation magnitude for MoE activation quantization.

REPRODUCIBILITY STATEMENT

Section 4 defines the calibration rule and fixed inference steps; Appendix H specifies the calibration construction and implementation details, while Appendix I documents the evaluation protocols. Protocols for the controls, ablations, and diagnostics are given in their respective appendix sections.

AI USE STATEMENT

Generative AI tools were used for Chinese-to-English translation, language polishing, and table formatting, and early idea related works retrieval. They were also used to assist with refining the motivation, designing selected experiments, including the matched-λ-Absmax control and the context-floor quantile ablation, implementation, result summarization, and parts of the literature search. All AI-assisted code, analyses, references, and manuscript content were manually reviewed and verified by the authors. All reported experimental results and scientific conclusions were independently validated by the authors, who take full responsibility for the final manuscript.

REFERENCES

Aida Amini, Saadia Gabriel, Shanchuan Lin, Rik Koncel-Kedziorski, Yejin Choi, and Hannaneh Hajishirzi. MathQA: Towards interpretable math word problem solving with operation-based formalisms. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 2357–2367, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1245. URL https://aclanthology.org/N19-1245/.

Guanjie Chen, Xinyu Zhao, Tianlong Chen, and Yu Cheng. MoE-RBench: Towards building reliable language models with sparse mixture-of-experts. In Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp (eds.), *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 7792–7808. PMLR, 21–27 Jul 2024. URL https://proceedings.mlr.press/v235/chen24bg.html.

Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde de Oliveira Pinto, Jared Kaplan, Harri Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, Alex Ray, Raul Puri, Gretchen Krueger, Michael Petrov, Heidy Khlaaf, Girish Sastry, Pamela Mishkin, Brooke Chan, Scott Gray, Nick Ryder, Mikhail Pavlov, Alethea Power, Lukasz Kaiser, Mohammad Bavarian, Clemens Winter, Philippe Tillet, Felipe Petroski Such, Dave Cummings, Matthias Plappert, Fotios Chantzis, Elizabeth Barnes, Ariel Herbert-Voss, William Hebgen Guss, Alex Nichol, Alex Paino, Nikolas Tezak, Jie Tang, Igor Babuschkin, Suchir Balaji, Shantanu Jain, William Saunders, Christopher Hesse, Andrew N. Carr, Jan Leike, Josh Achiam, Vedant Misra, Evan Morikawa, Alec Radford, Matthew Knight, Miles Brundage, Mira Murati, Katie Mayer, Peter Welinder, Bob McGrew, Dario Amodei, Sam McCandlish, Ilya Sutskever, and Wojciech Zaremba. Evaluating large language models trained on code. *arXiv preprint arXiv:2107.03374*, 2021. URL https://arxiv.org/abs/2107.03374.

Yuanteng Chen, Yuantian Shao, Peisong Wang, and Jian Cheng. EAC-MoE: Expert-selection aware compressor for mixture-of-experts large language models. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 12942–12963, Vienna, Austria, July 2025a. Association for Computational Linguistics. doi: 10.18653/v1/2025.acl-long.633. URL https://aclanthology.org/2025.acl-long.633/.

Zhixuan Chen, Xing Hu, Dawei Yang, Zukang Xu, Xu Chen, Zhihang Yuan, Sifan Zhou, and Jiangyong Yu. MoEQuant: Enhancing quantization for mixture-of-experts large language models via expert-balanced sampling and affinity guidance. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 8245–8260. PMLR, 2025b. URL https://proceedings.mlr.press/v267/chen25aa.html.

Mohammed Nowaz Rabbani Chowdhury, Kaoutar El Maghraoui, Hsinyu Tsai, Naigang Wang, Geoffrey W. Burr, Liu Liu, and Meng Wang. Efficient quantization of mixture-of-experts with theoretical generalization guarantees. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/7b97d09f0ab0f52eab856b0ed1122456-Abstract-Conference.html.

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

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*, 2021. URL https://arxiv.org/abs/2110.14168.

Elias Frantar, Saleh Ashkboos, Torsten Hoefler, and Dan Alistarh. OPTQ: Accurate quantization for generative pre-trained transformers. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=tcbBPnfwxS.

Zhongqian Fu, Tianyi Zhao, Ning Ding, Xianzhi Yu, Xiaosong Li, Yehui Tang, and Yunhe Wang. EAQuant: Enhancing post-training quantization for MoE models via expert-aware optimization. *arXiv preprint arXiv:2506.13329*, 2026. URL https://arxiv.org/abs/2506.13329v3.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=d7KBjmI3GmQ.

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. URL https://proceedings.mlsys.org/paper_files/paper/2024/hash/42a452cbafa9dd64e9ba4aa95cc1ef21-Abstract-Conference.html.

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

Marco Tulio Ribeiro, Tongshuang Wu, Carlos Guestrin, and Sameer Singh. Beyond accuracy: Behavioral testing of NLP models with CheckList. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault (eds.), *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 4902–4912, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.442. URL https://aclanthology.org/2020.acl-main.442/.

Freda Shi, Xinyun Chen, Kanishka Misra, Nathan Scales, David Dohan, Ed H. Chi, Nathanael Schärli, and Denny Zhou. Large language models can be easily distracted by irrelevant context. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett (eds.), *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 31210–31227. PMLR, 23–29 Jul 2023. URL https://proceedings.mlr.press/v202/shi23a.html.

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, volume 202 of *Proceedings of Machine Learning Research*, pp. 38087–38099. PMLR, 2023. URL https://proceedings.mlr.press/v202/xiao23c.html.

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, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1472. URL https://aclanthology.org/P19-1472/.

A CONTEXTUALIZED-EXECUTION SANITY CHECK

Experimental setup. We use the unquantized full-precision Qwen3.5-35B-A3B reference and sample 128 MMLU target examples A_i. Each target is evaluated alone and after each of eight unrelated natural-language prefixes B_j containing 87–309 tokens, giving 1,024 target–prefix pairs. We evaluate only the target example. Let $\hat{y}_i^A$ and $\hat{y}_{i,j}^{BA}$ denote the two predicted answers. We measure the answer preservation rate

$$\mathrm{APR} = \frac{1}{N_A N_B} \sum_i \sum_j \mathbf{1}[\hat{y}_i^A = \hat{y}_{i,j}^{BA}]$$

and the pre-router representation displacement

$$d_{i,j,l} = \|h_l(B_j A_i) - h_l(A_i)\|_2.$$

These diagnostics are used only to characterize the contextualized executions.

Results. Table 3 reports prediction preservation.

The original answer is preserved for 940 of the 1,024 pairs. Among pairs whose original answer is correct, 97.11% remain correct after prefixing.

The prefixes produce substantial representation changes despite the high answer preservation rate. For example, the median displacement at layer 24 is 32.79, compared with a median reference-state norm of 38.45. Preserved pairs do not consistently have smaller displacement than flipped pairs.

Unrelated prefixes preserve the model decision for most evaluated pairs while substantially changing internal representations, supporting their use as alternative calibration executions.

B MATCHED STEP-INFLATION CONTROL

Experimental setup. This control tests whether CR benefits only from increasing the quantization step relative to Absmax. We use Qwen3.5-35B-A3B, the default context aggregation quantile $p = 0.5$, and 1,024 paired evaluation samples. All methods use the same calibration artifacts, evaluation samples, and fake-quantization path. Across 20,480 quantization-site keys, 80 keys have no site-level instability signal and fall back to Absmax.

At $p = 0.5$, the CR floor determines 98.865% of the 26,214,400 step elements. The control therefore compares CR with three rescaled Absmax baselines. For the set of step elements $\mathcal{I}$, their coefficients are

$$\lambda_{\mathrm{mass}} = \frac{\sum_{(k,c)\in\mathcal{I}} \Delta_{\mathrm{CR},k,c}}{\sum_{(k,c)\in\mathcal{I}} \Delta_{\mathrm{abs},k,c}}, \qquad \lambda_{\mathrm{median}} = \mathrm{Median}_{(k,c)\in\mathcal{I}} \frac{\Delta_{\mathrm{CR},k,c}}{\Delta_{\mathrm{abs},k,c}},$$

$$\lambda_k = \frac{\sum_c \Delta_{\mathrm{CR},k,c}}{\sum_c \Delta_{\mathrm{abs},k,c}}, \qquad \Delta_{\lambda,k,c} = \lambda \Delta_{\mathrm{abs},k,c}.$$

The first two controls match the global step mass (sum of quantization steps) and median step ratio; the third matches step mass within each expert quantization site. We obtain $\lambda_{\mathrm{mass}} = 10.53$ and $\lambda_{\mathrm{median}} = 9.56$. The per-site coefficients have median 9.50, 90th percentile 19.5, and maximum 191.

For sample i, we report

$$\Delta\mathrm{PPL}_i = \mathrm{PPL}_{i,\mathrm{quant}} - \mathrm{PPL}_{i,\mathrm{ref}}.$$

Here ref denotes the unquantized full-precision reference.

Results. CR and Absmax have similar mean degradation: their paired mean difference is 0.0099 with 95% CI $[-0.0436, 0.0597]$. CR has a variance ratio of 0.373 $[0.216, 0.637]$ relative to Absmax, and reduces P99 from 4.61 to 2.67. Relative to the mass- and median-matched controls, CR's paired mean differences are -0.8182 $[-0.8756, -0.7634]$ and -0.6314 $[-0.6852, -0.5797]$, respectively. Per-site matching narrows the gap, but CR still improves mean ΔPPL by 0.1209 $[0.0868, 0.1552]$: step magnitude alone is insufficient; floor placement also matters.

Table 3: Prediction preservation under contextualized execution on Qwen3.5-35B-A3B. Brackets denote 95% confidence intervals.

Metric	A only	BA	Change / preservation
MMLU accuracy	0.8438	0.8496	$+0.0059\ [-0.0332, 0.0459]$
Answer preservation rate	–	–	$0.9180\ [0.8701, 0.9600]$
Target-level robust rate	–	–	$0.9062\ [0.8516, 0.9531]$
Originally-correct preservation	–	–	$0.9711\ [0.9369, 0.9977]$
Originally-incorrect correction rate	–	–	$0.1938\ [0.0435, 0.3850]$
Median $\Delta \log p$ of baseline winner	–	–	$-0.0054\ [-0.0101, -0.0014]$
Median answer margin	3.5625	3.5938	$+0.0781$

Table 4: Pre-router representation displacement $\|h_l(BA) - h_l(A)\|_2$ at the MMLU decision position.

Layer	Median	P90	Preserved med.	Flipped med.	Median $\|h_l(A)\|_2$
0	1.78	2.49	1.81	1.36	22.22
4	10.41	12.62	10.45	9.07	28.47
9	19.33	22.79	19.43	17.74	31.80
14	22.68	41.57	23.48	19.30	31.93
19	31.10	56.98	32.57	24.00	40.99
24	32.79	55.08	34.38	24.65	38.45
29	36.54	55.77	37.18	27.26	45.09
34	32.57	67.69	32.84	27.64	58.07
39	24.46	35.95	24.46	24.51	54.94

B.1 Sensitivity to Instability-to-Floor Mapping

Experimental setup. We keep the median site-level routed range-instability signal fixed and vary only the floor-strength coefficient:

$$\tau_k = Q_{0.5}(\{|R_{A,k,c} - R_{BA,k,c}|\}_c)\,, \qquad \Delta_{k,c}^{(\gamma)} := \max(\Delta_{\mathrm{abs},k,c}, \gamma\tau_k)\,.$$

We evaluate $\gamma \in \{0, 0.5, 1, 2\}$ on Qwen3.5-35B-A3B with FP16 weights and INT8 routed-expert activations. The evaluation uses 1,024 held-out SkyPile records, indices 1024–2047, with maximum sequence length 2048.

Results. $\gamma = 0.5$ gives the lowest mean degradation and P95, while the default $\gamma = 1$ gives the lowest variance, P99, maximum, and exceedance rates. $\gamma = 2$ worsens both average and tail degradation. Moderate floor values therefore reduce degradation dispersion, whereas an overly large floor coarsens the quantization grid.

C Same-Structure Non-Contextual Control

Experimental setup. This control tests whether the shared-floor structure alone explains CR. Let $\mathcal{D}_A$ be the original calibration collection and $\mathcal{D}_C$ an independently sampled calibration collection from the same source corpus. Both use the same collector and natural-routing protocol. We define

$$\tau_k^{\mathrm{NC}} = Q_{0.5}(\{|R_{A,k,c} - R_{C,k,c}|\}_c)\,, \qquad \Delta_{k,c}^{\mathrm{NC}} = \max\!\left(\Delta_{k,c}^{\mathrm{abs}}, \beta\tau_k^{\mathrm{NC}}\right).$$

The coefficient β matches the total step mass of CR:

$$\sum_{k,c} \Delta_{k,c}^{\mathrm{NC}} = \sum_{k,c} \Delta_{k,c}^{\mathrm{CR}}.$$

We evaluate two independent draws of $\mathcal{D}_C$, with $\beta = 0.506245$ and 0.542744. Their step-mass ratios to CR are 1.000000046 and 1.000000034. Qwen3.5-35B-A3B is evaluated under W16A8 on 1,024 held-out SkyPile records, indices 1024–2047, with maximum sequence length 2048. The full-precision reference, Absmax, CR, and both NC variants run in the same process.



Figure 3: Post-hoc t-SNE visualization of layer-24 MoE-entry vectors for target tokens without a prefix (diamonds) and under 32 prefixes (circles). The displayed groups were selected for visual clarity and serve as a qualitative illustration of how prefixes shift MoE-entry representations.

Table 5: Source of the final CR quantization step at each context aggregation quantile. "Absmax (eligible)" excludes keys without a site-level instability signal.

p	Absmax (all)	Absmax (eligible)	All Absmax	All context	Mixed
0.10	20.897%	20.587%	751	1,126	18,521
0.50	1.132%	0.745%	0	15,894	4,506
0.90	0.401%	0.0103%	0	19,990	410
0.99	0.391%	0.000023%	0	20,397	3

Results. The two NC draws have mean ΔPPL values of 2.2875 and 2.4024, compared with 0.5013 for CR, and are worse on 1008 and 1006 of the 1,024 samples, respectively. Their degradation variances are $7.71\times$ and $8.59\times$ that of CR. Despite matched total step mass, 99.5% of individual step elements differ between NC and CR: shared-floor structure and step mass are insufficient without contextual site-wise placement.

D Context Aggregation Quantile Ablation

Experimental setup. For expert quantization site k and channel c, we define the channelwise routed-range discrepancy

$$g_{k,c} = |R_{A,k,c} - R_{BA,k,c}|, \qquad \tau_k(p) := Q_p(\{g_{k,c}\}_c)\,,$$

$$\Delta^{\mathrm{CR}}_{k,c}(p) = \max\left(\frac{M_{k,c}}{127}, \tau_k(p)\right),$$

where $M_{k,c}$ is the Absmax statistic. The main experiments use $p = 0.5$.

D.1 Per-Sample Perplexity Sensitivity

We evaluate $p \in \{0.10, 0.50, 0.90, 0.99\}$ using Qwen3.5-35B-A3B with FP16 weights and INT8 routed-expert activations. The four settings use the same 1,024 held-out SkyPile records, indices 1024–2047, maximum sequence length 2048, and the same full-precision reference.

Results. $p = 0.10$ lowers mean ΔPPL by 0.1163 relative to $p = 0.50$, but $p = 0.50$ has lower variance, P95, and P99. The mean step rises from 0.0632 at $p = 0.10$ to 0.3023, 0.7999, and 1.5736

Table 6: Sample-level perplexity degradation for CR and matched λ-Absmax step controls. Lower values are better.

Method	Mean	Variance	p99	Max	$\Pr(\Delta > 1)$	$\Pr(\Delta > 5)$
Absmax	0.4870	1.1355	4.61	17.87	13.0%	0.98%
CR	0.4969	0.4232	2.67	12.11	10.9%	0.10%
λ_k-Absmax	0.6178	0.5521	3.13	11.02	17.4%	0.20%
λ_{median}-Absmax	1.1283	1.0660	5.01	10.06	43.3%	1.07%
λ_{mass}-Absmax	1.3151	1.3713	5.64	12.29	52.5%	1.76%

Table 7: Sensitivity to the instability-to-floor coefficient on Qwen3.5-35B-A3B W16A8 with fixed $p = 0.5$.

γ	Mean ΔPPL	Variance	P95	P99	Max	$\Pr(\Delta > 1)$	$\Pr(\Delta > 5)$
0 (Absmax)	0.4879	1.1355	1.9027	4.6567	17.8677	0.1289	0.0098
0.5	**0.4554**	0.4221	**1.3106**	2.4708	12.7860	0.1074	0.0020
1	0.4843	**0.4019**	1.3580	**2.4066**	**12.1105**	**0.1055**	**0.0010**
2	1.6894	1.9051	4.2411	7.0910	12.6379	0.6689	0.0381

at the three larger quantiles. At $p \geq 0.90$, nearly all step elements use the CR floor and both average and tail degradation increase sharply. The quantile therefore controls a range–resolution trade-off rather than providing monotonic improvement.

D.2 Downstream Check of the Context Aggregation Quantile

Experimental setup. We compare $p = 0.10$ and $p = 0.50$ under W16A8 on fixed, paired subsets. MMLU uses four examples from each of 57 subjects, giving 228 questions, with 5-shot evaluation and seed 0. BoolQ uses 256 examples. The subsets are selected with seed 17 and both quantiles use the same items in the same run.

Results. $p = 0.10$ is 1.75 percentage points below $p = 0.50$ on MMLU and 3.52 points above it on BoolQ, with comparable macro averages on the paired subsets. We use $p = 0.50$ as the default because it has the best tail behavior among the tested moderate settings while maintaining comparable downstream performance.

E Routed-Count-Matched Contextual Range Diagnostic

Experimental setup. This diagnostic tests whether different routed observation counts explain the channelwise contextual routed-range discrepancy

$$g_{k,c} = |R_{A,k,c} - R_{BA,k,c}|.$$

We use Qwen3.5-35B-A3B and a paired 1:1 protocol. At each routed quantization site, the original and contextualized paths are subsampled to equal counts before computing $g^{\text{ctx}}_{k,c}$. We compare it with two references: a conservative null that splits the original collection into disjoint matched subsets, and an independent-sampling null that draws two matched subsets from the original collection with overlap allowed. The site-level routed range-instability signal is

$$\tau_k = Q_{0.5}(\{g_{k,c}\}_c)\,.$$

The median routed counts before matching are 1,382 and 1,453, with a median ratio of 0.975. After routed-count and minimum-support filtering, 20,026 of 20,472 sites remain. All subsampling experiments use five seeds.

Results. The contextual discrepancy is 0.8064 times the conservative split null, so it does not exceed all within-collection range variation. It is 2.2847 times the independent-sampling null, and the corresponding site-level τ_k ratio is 2.2373. These values are stable across the five seeds. Compared with the historical approximately 1:32 collection, matching counts reduces median τ_k from 0.2031 to

Table 8: Same-structure non-contextual control on Qwen3.5-35B-A3B W16A8. All quantized rows report sample-level degradation relative to FP16.

Method	Corpus PPL	Mean ΔPPL	95% CI	Median	Variance	P99	$\Pr(\Delta > 1)$	$\Pr(\Delta > 5)$
Full-precision	12.5824	0	–	0	0	0	0	0
Absmax	12.9637	0.4872	[0.4269, 0.5546]	0.2026	1.0642	4.578	0.136	0.0098
CR	13.0262	0.5013	[0.4585, 0.5493]	0.3706	0.5336	2.728	0.102	0.0020
NC	14.5289	2.2875	[2.1677, 2.4141]	1.7436	4.1131	10.737	0.772	0.0732
NC-draw2	14.5954	2.4024	[2.2746, 2.5358]	1.8270	4.5865	10.842	0.787	0.0781

Table 9: Paired comparisons for the same-structure control. Differences are $\Delta\text{PPL}^{\text{alternative}} - \Delta\text{PPL}^{\text{baseline}}$.

Comparison	Mean difference	95% CI	Samples worse
NC − CR	+1.7862	[1.6883, 1.8900]	1008 / 1024
NC-draw2 − CR	+1.9011	[1.7941, 2.0133]	1006 / 1024
Absmax − CR	−0.0141	[−0.0655, 0.0410]	–
NC-draw2 − NC	+0.1148	[0.0691, 0.1620]	–

0.172607 and its ratio to the median Absmax step from about $8.3\times$ to $7.0\times$. Observation-count imbalance contributes to τ_k, while the contextual discrepancy remains above the independent-sampling null. These results characterize τ_k as a context-informed routed-range instability statistic containing both contextual and finite-sample variation.

F Absmax Sensitivity to Calibration Collection Sampling

Experimental design. We test whether the HumanEval degradation of W16A8 Absmax can be removed by changing its calibration samples while holding the Qwen3.5-35B-A3B FP16 weights, expert-input INT8 fake quantizer, generation settings, and evaluation items fixed. Four calibration collections each contain 1,024 disjoint random records from the same SkyPile shard; the first 2,048 raw lines are excluded and duplicate texts are rejected. A fifth collection contains 256 additional disjoint SkyPile records and 128 training records each from MMLU, BoolQ, HellaSwag, OpenBookQA, MathQA, and GSM8K, for a total of 1,024. The non-SkyPile records are formatted as questions or prompts followed by their correct answers. HumanEval is not used for calibration. All five collections use seed 20260924 to freeze the draw, a 2,048-token maximum length, and all-token, naturally routed Absmax range collection.

From the 164 official HumanEval tasks, archived per-task outcomes identify 78 historical FP16-pass/Absmax-fail tasks; we select 32 with seed 20260924. All five new step tensors are evaluated on these same 32 tasks with greedy generation (at most 512 new tokens) and the official functional tests. The FP16, Absmax, and CR rows use outcomes from archived full-164 runs restricted to this subset; CR uses its original calibration.

Results and interpretation. Table 14 shows a 3–9/32 recovery range across the four SkyPile draws: calibration sampling changes which failures recover, but 22 of the 32 tasks fail under all four draws. The mixed collection recovers three tasks that none of the four SkyPile-only collections recovers; across all five collections, 19/32 tasks always fail. Mixed reaches 9/32, equal to the best SkyPile-only draw. Neither resampling SkyPile nor adding the six other datasets consistently recovers the selected failures. The mixed collection contains 249,735 effective tokens versus 546,575–592,790 for SkyPile-only draws. This experiment tests calibration-sample sensitivity on the selected failure subset; the historical CR row is included as a reference.

G Detailed HumanEval Autoregressive Generation-Drift Analysis

Experimental setup. We compare Absmax, EBSS, and CR under W16A8 on all 164 HumanEval prompts using the same FP16 weight checkpoint. The analysis covers 9,877 full-precision reference

Table 10: Per-sample perplexity degradation under different context aggregation quantiles. "CR-floor-selected" is the fraction of step elements for which the CR floor is selected.

p	CR-floor-selected	Mean ΔPPL	Median	Variance	P95	P99	$\Pr(\Delta > 1)$
0.10	79.103%	0.3720	0.1569	0.7444	1.5256	3.2269	0.0947
0.50	98.868%	0.4883	0.3596	0.4341	1.3597	2.8751	0.1074
0.90	99.599%	4.3708	3.5410	11.5797	11.0898	16.3322	0.9580
0.99	99.609%	14.9379	12.2799	121.4605	34.9987	55.8218	0.9971

Table 11: Paired comparison with the default $p = 0.50$. The difference is $\Delta\mathrm{PPL}(p) - \Delta\mathrm{PPL}(0.50)$ over 10,000 bootstrap resamples.

Comparison	Mean difference	95% CI	Better / worse samples	Variance ratio
$p = 0.10$ vs. 0.50	-0.1163	$[-0.1555, -0.0748]$	710/314	1.715
$p = 0.90$ vs. 0.50	$+3.8825$	$[+3.7040, +4.0734]$	3/1021	26.68
$p = 0.99$ vs. 0.50	$+14.4496$	$[+13.8172, +15.1137]$	0/1024	279.8

continuation tokens. Teacher-forced evaluation measures prediction changes under the same reference token history, while short greedy decoding measures sequence divergence. We also collect clipping rate, relative RMSE, and clipping-excess RMS at the gate/up and down-projection inputs.

Results. CR reduces top-1 disagreement to 2.92% and mean ΔNLL to 0.01929, compared with 8.36% and 0.18715 for Absmax and 9.04% and 0.21161 for EBSS. The first teacher-forced divergence moves from about 6 tokens to 15.5 tokens, and CR exactly matches the reference greedy continuation on 37 of 164 prompts, compared with 4 for Absmax and 3 for EBSS.

CR does not minimize local reconstruction error at every site: its gate/up relative RMSE is 0.13092, above both baselines. It nevertheless reduces the down-input clipping rate from about 0.24% to 0.01480% and has the lowest clipping-excess RMS at both sites. CR therefore reduces autoregressive generation drift mainly by reducing rare clipping and early prediction changes, rather than by uniformly reducing local quantization error.

H Algorithm and Implementation Details

H.1 Nominal Calibration Algorithm

Table 18 specifies quantization-step construction when all three statistics are available. An observation mask selects which tokens contribute to calibration statistics; it does not force expert execution. Thus prefix tokens may participate in the forward computation even if excluded from statistic collection.

H.2 Models and Quantization Configuration

We use the base checkpoints https://huggingface.co/Qwen/Qwen3.5-35B-A3B and https://huggingface.co/google/gemma-4-26B-A4B. W16A8 retains the reference weights; W4A8 uses AWQ, GPTQ, or SmoothQuant-based (SQ) weight transforms. For the CR and Absmax AWQ/GPTQ configurations, weight calibration reads the first 256 SkyPile examples. SQ instead computes smoothing factors from activation ranges. EBSS is an external baseline.

Activation calibration targets the routed experts' gate/up and down input sites. The range collector stores channelwise absolute maxima on CPU in FP32. At $p = 0.5$, the site-level discrepancy quantile uses linear interpolation (the mean of the two middle ordered values for even channel counts). The main CR configuration uses floor coefficient $\gamma = 1$. The A8 simulation divides by fixed steps in FP32, rounds to the nearest integer, clips to $[-128, 127]$, dequantizes, and converts back to the activation dtype. No calibration prefix or scale adaptation is required at inference.

Table 12: Downstream comparison on fixed evaluation subsets. ΔAcc is accuracy$(p = 0.10)$ minus accuracy$(p = 0.50)$.

Task	$p = 0.10$	$p = 0.50$	ΔAcc	95% CI	.10 wins	.50 wins
MMLU-228	0.8333	0.8509	-0.0175	$[-0.0482, +0.0132]$	4	8
BoolQ-256	0.8984	0.8633	$+0.0352$	$[+0.0000, +0.0703]$	15	6
Macro average	0.8659	0.8571	$+0.0088$	$[-0.0136, +0.0318]$	–	–

Table 13: Routed-count-matched contextual range results. Values are medians over five subsampling seeds; brackets give seed-wise IQRs.

Metric	Conservative split null	Independent-sampling null
$\mathrm{med}(g^{\mathrm{ctx}})$	0.226807	0.226807
Reference median	0.281250	0.099487
g^{ctx}/reference	0.8064 [0.8056, 0.8073]	2.2847 [2.2770, 2.2857]
$\Pr(g^{\mathrm{ctx}} >$ reference$)$	0.4439	0.6567
$\mathrm{med}_k\, \tau_k^{\mathrm{ctx}}$	0.218750	0.218750
Reference $\mathrm{med}_k\, \tau_k$	0.283203	0.101562
τ^{ctx}/reference	0.8000 [0.7989, 0.8000]	2.2373 [2.2373, 2.2469]

Table 14: Qwen3.5-35B-A3B HumanEval recovery under W16A8 Absmax calibration resampling on 32 historically identified FP16-pass/Absmax-fail tasks. Calibration token counts follow 2,048-token truncation; FP16, Absmax, and CR are historical references, and the five collection rows are new measurements.

Calibration collection	Calibration tokens	Passed / 32	Pass rate (%)
Historical FP16 (W16A16)	—	32	100.0
Historical W16A8 Absmax	—	0	0.0
SkyPile draw 1 (1,024)	592,790	9	28.1
SkyPile draw 2 (1,024)	546,575	3	9.4
SkyPile draw 3 (1,024)	583,166	6	18.8
SkyPile draw 4 (1,024)	589,731	4	12.5
Mixed (256 SkyPile $+6 \times 128$)	249,735	9	28.1
Historical W16A8 CR	—	24	75.0

H.3 Calibration examples and token observation

The unprefixed CR collection uses the first 32 SkyPile targets. For its contextualized collection, the first 32 GSM8K training examples are each paired with the first 32 SkyPile examples, yielding $32 \times 32 = 1{,}024$ prefixed texts. Each text has the form `Question: {question} Answer: {answer} News: {text}`. The stored prefix length counts characters through the GSM8K answer but excludes the `News:` separator. The independent Absmax range collection uses the first 1,024 SkyPile examples. The 32 unprefixed CR targets are therefore contained in the Absmax collection.

Prefix tokens still execute and affect routing and hidden states. The collector can observe all tokens (mode 0) or start observation at the tokenized character-prefix length (mode 1); truncation is at 2,048 tokens.

H.4 Safeguard and calibration coverage

At a site with both contextual statistics and an Absmax range, the final step is $\max(M_{k,c}/127, \tau_k, 10^{-8})$; a site with only Absmax uses the Absmax step. Under the current runner, an expert is activation-quantized only if both gate/up and down site steps are available; otherwise it remains at reference activation precision.

Table 15: Prediction drift on HumanEval under W16A8 expert activation quantization. Top-1 disagreement and ΔNLL use the reference token history.

Method	Top-1 disagree. (%)	Mean ΔNLL	First teacher div.	Greedy div.	Exact match
Absmax	8.36	0.18715	6.3	160 / 164	4 / 164
EBSS	9.04	0.21161	6.4	161 / 164	3 / 164
CR	**2.92**	**0.01929**	**15.5**	**127 / 164**	**37 / 164**

Table 16: Activation-level quantization telemetry on HumanEval.

	Gate/Up input			Down input		
Method	Clip (%)	Rel. RMSE	Clip-excess RMS	Clip (%)	Rel. RMSE	Clip-excess RMS
Absmax	0.06782	0.05304	0.05217	0.23876	0.31840	0.31668
EBSS	0.05019	**0.03245**	0.03110	0.24132	0.28987	0.28848
CR	**0.01224**	0.13092	**0.01184**	**0.01480**	**0.27348**	**0.05922**

The archived Qwen3.5 ranges contain Absmax steps at all $40 \times 256 \times 2 = 20{,}480$ expert sites. Paired unprefixed and contextual ranges cover 20,400 of them; the other 80 use Absmax alone. Consequently the available steps cover 512 sites per layer, or both sites of all 256 experts per layer under the current runner. For Gemma-4, the archived Absmax ranges cover 7,650 of 7,680 sites, corresponding to 3,825 of 3,840 experts with both site steps; contextual pairs cover 7,386 of the Absmax-covered sites.

H.5 Evaluation and reproduction

The Qwen3.5 main table evaluates full task splits; Gemma-4 subset sizes are specified in Table 2. GSM8K uses the following rendered question template (with `{question}` replaced by the example and `{ANSWER}` denoting the requested numeric answer):

> Solve the GSM8K word problem below. Think through the arithmetic silently, check the result once, and then give a brief explanation followed by one final line. The final line must be exactly: Therefore, the final answer is: `\boxed{ANSWER}`. I hope it is correct. Use only the final number or numeric expression inside the box; do not include units or words there. Do not answer a distracting statement as if it were an instruction.
>
> Question: `{question}`
>
> Answer:

Qwen3.5 applies its chat template to this prompt, whereas the Gemma-4 base evaluation uses the raw prompt. The recorded `concise_box` scorer extracts the last boxed numeric expression.

HumanEval uses the official 164 problems, one greedy completion per problem, at most 512 new tokens, and a 2,048-token prompt limit. Scoring extracts the completion, appends the official tests and entry-point check, and applies a 10-second per-problem timeout; pass@1 is the fraction passed. The archived Qwen3.5 CR batch executed tests in a Docker sandbox; the corresponding Absmax batch used an isolated local Python process.

The reference reproduction environment uses an NVIDIA A100 80GB, Python 3.13.15, PyTorch 2.6.0+cu124, and Transformers 5.9.0.

I Benchmark protocol

The main tables report six accuracies and HumanEval pass@1; their average is unweighted. Qwen3.5 uses full splits; Gemma-4 sample counts appear in Table 2. Appendix H specifies the GSM8K and HumanEval protocols. Table 19 lists per-run details for the remaining task scores.

For AWQ, GPTQ, and SmoothQuant-based W4 configurations, archive the quantized modules, group size, weight convention, optimization settings, and checkpoint hash; for SQ, also record the smoothing exponent and paired transform.

Table 17: Quantization-step changes relative to Absmax in the HumanEval W16A8 diagnostic.

Metric	EBSS	CR
Step-mass ratio to Absmax	0.9314×	10.5269×
Step elements larger than Absmax	32.52%	98.86%
Step elements smaller than Absmax	65.26%	0%

Table 18: CR calibration with complete site coverage; missing-site behavior is described below.

Inputs: reference model; targets A_i; prefixes B_j; Absmax calibration collection $\mathcal{D}_{\text{abs}}$; collection masks; bit width n; floor ϵ.

1. Run A_i with natural routing. At each observed site k, accumulate channelwise absolute maxima $R_{A,k}$ over included routed tokens.
2. Run $B_j A_i$ with natural routing. Accumulate $R_{BA,k}$ over the included routed observations and all contextual passes. Do not assume paired expert assignments.
3. Run $\mathcal{D}_{\text{abs}}$ and accumulate independent channelwise maxima M_k at the same quantization sites.
4. For every site with both contextual statistics, compute $\tau_k = Q_{0.5}(\{|R_{A,k,c} - R_{BA,k,c}|\}_c)$.
5. For sites with all statistics, set $\Delta_{k,c} = \max(M_{k,c}/q_{\max}^{(n)}, \tau_k, \epsilon)$. Record incomplete sites separately.
6. Save fixed steps, site coverage, and the calibration manifest. At inference, apply quantize–dequantize at the gate/up and down inputs after dispatch.

J Routing Diagnostics and Discussion

Because activation quantization is applied after expert dispatch, it does not change the routing decision of the current MoE layer, but its perturbation can propagate to later layers and affect subsequent expert selection. We therefore inspect routing changes as a possible intermediate effect of expert activation quantization.

Table 20 summarizes the available routing diagnostics. In the aligned export, the reference and quantized executions disagree on their top-8 expert sets for 38.21% of token–layer positions, while the average reference routing mass assigned to removed experts is only 3.90%. Conditioned on a set disagreement, the removed mass rises to 10.21%. This suggests that expert-set disagreement and the magnitude of the routing change capture different aspects of the perturbation.

The site ablation also suggests that perturbations introduced at different expert activation sites have different effects on later routing. Quantizing the down-projection input produces substantially more set disagreement than quantizing only the gate/up input, while quantizing both sites produces only a small additional increase. The sparse depth-wise measurements further suggest that routing sensitivity can vary across layers.

These observations motivate several questions for future study. In particular, it would be useful to examine whether routing changes are associated with hidden-state drift and output degradation,

Table 19: Evaluation settings to archive for each main-table run.

Task	Required run-specific protocol
MMLU	Split, subject aggregation, shots, prompt, answer scoring/normalization.
BoolQ	Split, verbalizers, prompt, likelihood versus generation.
HellaSwag	Split, preprocessing, completion scoring/normalization.
OpenBookQA	Split, supporting facts, choice formatting/scoring.
MathQA	Split, choice versus generation, prompt, answer parsing.
GSM8K	Prompt in Appendix H; record generation stops, token limit, numeric scorer version.
HumanEval	Official tests and greedy decoding; retain completion extraction, sandbox mode, timeout, and 164-task outputs per run.

Table 20: Available routing diagnostics under expert activation quantization.

Metric	Value (%)
Top-8 set disagreement	38.2055
Mean removed reference top-8 mass	3.9012
Removed mass given set disagreement	10.2112
Reference top-1 absent from quantized top-8	0.1918
Gate/up input only: set disagreement	23.3559
Down input only: set disagreement	40.4760
Both sites: set disagreement	40.5806
Layer 1: removed mass	1.2511
Layer 24: removed mass	5.4161
Layer 39: removed mass	3.6752

and whether routing probability mass provides a more informative signal than binary expert-set disagreement. A matched comparison among FP16, Absmax, and CR could further test whether the selective preservation observed for CR is accompanied by different patterns of later-layer routing drift.