FAQ: Mitigating Quantization Error via Regenerating Calibration Data with Family-Aware Quantization

Haiyang Xiao, Weiqing Li, Jinyue Guo, Guochao Jiang, Guohua Liu, Yuewei Zhang†

Alibaba Cloud Computing
{xiaohaiyang.xhy, liyou.zyw}@alibaba-inc.com

Abstract

Although post-training quantization (PTQ) provides an efficient numerical compression scheme for deploying large language models (LLMs) on resource-constrained devices, the representativeness and universality of calibration data remain a core bottleneck in determining the accuracy of quantization parameters. Traditional PTQ methods typically rely on limited samples, making it difficult to capture the activation distribution during the inference phase, leading to biases in quantization parameters. To address this, we propose **FAQ** (Family-Aware Quantization), a calibration data regeneration framework that leverages prior knowledge from LLMs of the same family to generate high-fidelity calibration samples. Specifically, FAQ first inputs the original calibration samples into a larger LLM from the same family as the target model, regenerating a series of high-fidelity calibration data using a highly consistent knowledge system. Subsequently, this data, carrying Chain-of-Thought reasoning and conforming to the expected activation distribution, undergoes group competition under expert guidance to select the best samples, which are then re-normalized to enhance the effectiveness of standard PTQ. Experiments on multiple model series, including Qwen3-8B, show that FAQ reduces accuracy loss by up to 28.5% compared to the baseline with original calibration data, demonstrating its powerful potential and contribution.

1 Introduction

The substantial computational and memory demands of large language models (LLMs) continue to drive inference optimization across algorithmic approaches and model-level techniques (Jiang et al., 2025; Quan et al., 2025), with weight quantization playing a central role. Post-training quantization (PTQ) converts pre-trained weights and activations from high-precision formats (e.g., FP16) to low-bit integers (e.g., INT8/INT4) without retraining, substantially reducing memory footprints and accelerating inference on resource-constrained hardware.

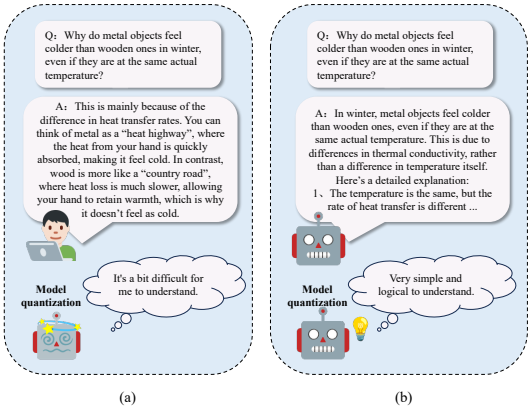


Figure 1: The impact of calibration data on quantization: (a) Traditional PTQ relies on human-provided calibration data, which may not align well with the model's internal activation patterns, leading to suboptimal quantization. (b) FAQ leverages a larger in-family model to generate a 'model-friendly' calibration set, ensuring better alignment and mitigating quantization errors.

Activation-distribution drift and quantization noise jointly degrade PTQ performance. Distribution drift arises as inputs vary across layers and during inference, causing mismatches between the quantization range (scale/zero-point) and the actual activations; this leads to accuracy loss. Moreover, quantization errors can propagate through nonlinearities, skip connections, and residuals, amplifying downstream degradation. Prior work mitigates these issues from multiple angles, including error-aware weight reconstruction (e.g., GPTQ (Frantar et al., 2023)), activation smoothing or scaling to tame outliers and range mismatch (e.g., SmoothQuant (Xiao et al., 2023), AWQ (Lin et al., 2024)), and lightweight calibration-time refinement/correction with a few iterations to suppress accumulated errors (e.g., OmniQuant (Shao et al.,

†Corresponding author.

Findings of the Association for Computational Linguistics: ACL 2026, pages 21441–21460
July 2-7, 2026 ©2026 Association for Computational Linguistics

2024), FPTQuant (van Breugel et al., 2025)).

However, many approaches fixate on the model level while neglecting calibration-sample adaptability. Calibration data may be constrained by privacy, and sample selection often requires multi-step optimization, introducing uncertainty. Real data closely tied to the target model may still fail to capture the complex internal activations during inference—especially in LLMs with prevalent outliers—leading to biased quantization parameters and notable accuracy degradation.

To address these challenges, we introduce **FAQ**, a **Q**uantization framework that optimizes calibration data by exploiting **F**amily **A**wareness. Models from the same development lineage and trained under similar paradigms exhibit substantial activation agreement; by reusing calibration mappings from "senior" family members, original samples with data bias can approximate the target model's activations, achieving distribution alignment under blind conditions. The resulting calibrated samples are refined via inter-group competition and content normalization to form a high-quality calibration set, thereby mitigating PTQ-induced errors.

Our contributions are as follows:

- We propose FAQ, a framework that leverages intra-family activation consistency to optimize calibration samples. To our knowledge, this is the first work to exploit family priors in PTQ calibration data.
- We show that family priors can be more influential for quantization performance than broader architectural similarities.
- Our pipeline uses chain-of-thought (CoT) guidance and intra-group competition to constrain calibration-data optimization, enabling seamless integration of FAQ with traditional PTQ methods.
- Extensive empirical comparisons with state-of-the-art (SOTA) PTQ baselines show that FAQ reduces quantization-induced accuracy loss by up to 28.5%, evidencing superior performance.

2 Related Work

2.1 Post-Training Quantization (PTQ)

Post-Training Quantization (PTQ) offers a low-cost, training-free solution for model compression but suffers significant accuracy degradation at low bit-widths, especially for LLMs (Choukroun et al., 2019; Hubara et al., 2020; Li et al., 2021a). Research to address this challenge has largely followed two primary directions: innovations in the quantization algorithms and data-driven optimizations of the calibration set.

Algorithm-level innovations have predominantly focused on mitigating quantization errors. One major line of work involves mixed-precision schemes, where sensitive outlier weights are stored at higher precision while non-critical values are aggressively quantized (Dettmers et al., 2024; Lee et al., 2024; Ou et al., 2024; Kim et al., 2024). Another strategy is **activation-aware quantization**, which identifies and protects weights that are multiplied by large activation values, as they are more critical to model performance (Lin et al., 2024; Huang et al., 2025). Others have focused on designing **advanced rounding mechanisms** or novel quantizers, such as additive quantization, to move beyond simple round-to-nearest schemes (Lee et al., 2023; Chee et al., 2023; Egiazarian et al., 2024). While powerful, these methods primarily treat the model's activation distribution as a fixed target to adapt to, rather than as a variable that can be optimized.

Data-driven approaches recognize the pivotal role of the calibration set. Li et al. (2021b) leveraged this data to perform layer-wise error minimization or block reconstruction. More recently, some works have focused on synthesizing calibration data when real data is inaccessible, for instance, Cai et al. (2020) proposed synthesizing data by matching batch normalization statistics. Li et al. (2026) proposes a method for measuring and synthesizing diverse data in an interpretable feature space. Gunasekar et al. (2023) demonstrated that small, high-quality synthetic datasets can outperform massive noisy corpora. To generate such high-quality data, Xu et al. (2024) introduced Evol-Instruct, a method that iteratively rewrites prompts to increase their complexity and diversity. Yu et al. (2025) further leverages Chain-of-Thought reasoning to generate high-quality samples. Our work, FAQ, advances this data-centric view. Instead of synthesizing data from statistical priors or merely using existing data to minimize local errors, we propose to regenerate a higher-quality calibration set from a more capable, in-family model, directly targeting the optimization of the activation distributions themselves.

2.2 Quantization-Aware Training (QAT)

Our approach should be distinguished from Quantization-Aware Training (QAT), which simulates quantization during a full fine-tuning phase (Jacob et al., 2018; Esser et al., 2020). While QAT can achieve superior accuracy, its high computational and data requirements are often prohibitive for LLMs. Our work remains strictly within the efficient, training-free PTQ paradigm, aiming to close the performance gap with QAT through data-centric innovations alone.

3 Theoretical Framework

3.1 Motivation

The fundamental goal of PTQ calibration is not to evaluate downstream task performance, which would require ground-truth labels, but rather to collect a small yet representative set of activations for the accurate computation of quantization parameters (e.g., scales and zero-points). From this standpoint, the critical factor is not the semantic fidelity of the calibration data to ground-truth, but the extent to which it can emulate the activation distributions the model encounters during inference.

Inspired by recent advanced methods such as GPTAQ (Li et al., 2025), we have found a key limitation of traditional quantizers like GPTQ (Frantar et al., 2023): their symmetric calibration objective fails to account for the fact that the input activations $\mathbf{A}$ are themselves the quantized outputs of the preceding layers. Specifically, the quantization objective can be formulated as:

$$O(W, \mathbf{A}) = \min_{Q(\cdot)\in\mathcal{Q}} \|Q(W)\mathbf{A} - W\mathbf{A}\|_F^2, \quad (1)$$

where W denotes the weight matrix, $\mathbf{A}$ represents the input activations, $Q(\cdot)$ is a quantization operator that maps W to a lower-precision space, and $\mathcal{Q}$ denotes the set of all feasible quantized weights. However, in practice, $\mathbf{A}$ is not the original, full-precision activation, but rather the quantized output from the preceding layer, introducing a mismatch between the calibration objective and the actual inference process. This discrepancy, where the calibration input differs from the true full-precision activations $\tilde{\mathbf{A}}$, becomes more pronounced in deeper layers. Therefore, a prevalent asymmetric calibration framework (Frantar et al., 2023) at this stage aims to correct for the propagated quantization error, which is formulated as:

$$O(W, \mathbf{A}, \tilde{\mathbf{A}}) = \min_{Q(\cdot)\in\mathcal{Q}} \|Q(W)\mathbf{A} - W\tilde{\mathbf{A}}\|_F^2. \quad (2)$$

While these algorithmic-level corrections are effective at managing challenging activation patterns, they are fundamentally compensatory in nature. The underlying issue is that the initial, full-precision activations $\tilde{\mathbf{A}}$ derived from standard calibration data are often inherently "hostile" to quantization, characterized by sparse, high-magnitude outliers. Our approach, FAQ, operates on a different and more fundamental premise. Instead of designing a more complex optimization objective to compensate for challenging activation patterns, we aim to proactively reshape the activation patterns themselves to be inherently more quantization-friendly.

3.2 Alignment of calibration data

Compared to directly modifying the quantization algorithm, FAQ induces a superior calibration set $\mathcal{D}_{\text{FAQ}}$ to elicit a new set of full-precision activations, denoted as $\widehat{\mathbf{A}}$.

$$O(W_n, \widehat{\mathbf{A}}_n) = \min_{Q} \|Q(W_n)\widehat{\mathbf{A}}_n - W_n\widehat{\mathbf{A}}_n\|_F^2 \quad (3)$$

$$O(W_n, \widehat{\mathbf{A}}_n, \tilde{\mathbf{A}}_n) = \min_{Q} \|Q(W_n)\widehat{\mathbf{A}}_n - W_n\tilde{\mathbf{A}}_n\|_F^2 \quad (4)$$

$$\widehat{\mathbf{A}}_{n+1} = Q(W_n)\widehat{\mathbf{A}}_n, \tilde{\mathbf{A}}_{n+1} = W_n\tilde{\mathbf{A}}_n, \tilde{\mathbf{A}}_0 = \widehat{\mathbf{A}}_0 \quad (5)$$

The effectiveness of FAQ stems from its ability to generate data that elicits more quantization-friendly activation patterns. When processing data $\widehat{\mathbf{A}}$ generated by FAQ, the activation distribution of a given layer is significantly smoother and more concentrated than the spiky and sparse distribution induced by the original calibration data $\tilde{\mathbf{A}}$, as visualized in Figure 2. By taming outliers at the data-source level, the quantization objective $O(W_n, \mathbf{A}_n, \widehat{\mathbf{A}}_n)$ becomes fundamentally easier to optimize. Consequently, even a standard, symmetric quantizer applied to our data can achieve superior performance, as described in Equation (3), Equation (4) and Equation (5). This is because the need for complex, asymmetric error correction is greatly diminished from the outset, making the subsequent quantization process more robust.

4 Main Method

We present FAQ, illustrated in Figure 1. The central idea of FAQ is to mitigate distributional mismatch between generic calibration data and a target

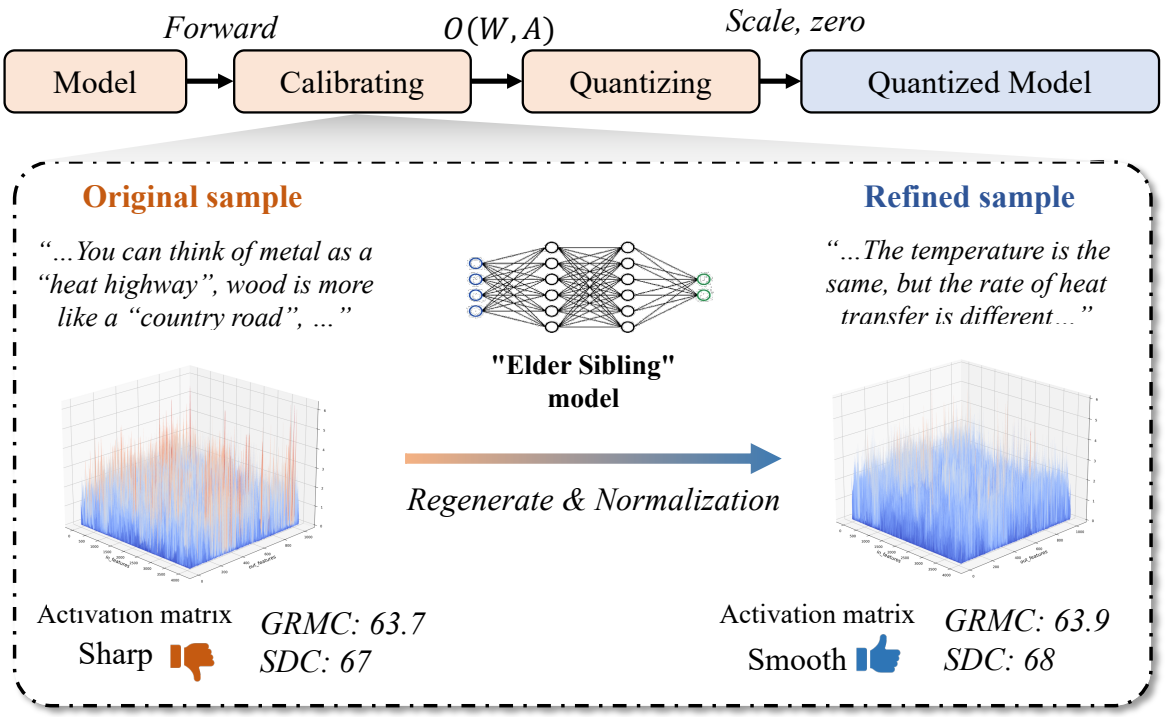


Figure 2: FAQ-enhanced PTQ calibration. Top: standard PTQ workflow. Bottom: zoom-in of the calibration stage. FAQ queries a larger in-family teacher ("elder-sibling") model (Qwen3-235B-A22B) to regenerate and normalize the original calibration prompts, yielding refined calibration data. The refined set induces smoother activation statistics in the target model (Qwen3-8B), illustrated by fewer extreme peaks in the activations, and improves PTQ results on both GRMC and SDC. GRMC (General Reasoning and Multilingual Capabilities) is the average score over 12 general downstream tasks; SDC (Specialized Domain Capabilities: Math and Code) is the average accuracy over AIME, MATH-500, and LiveCodeBench.

model's activation patterns. To achieve this, we replace the standard, pre-existing calibration set with a high-quality synthetic dataset specifically aligned with the target model's intrinsic characteristics.

4.1 Key Terminology

Before detailing our methodology, we clarify three key terms used throughout this paper:

- **Family**: Models sharing the same tokenizer, chat template, training paradigm, and alignment methodology within a common pre-training and post-training pipeline.
- **Regenerate**: The process by which an in-family "elder-sibling" dynamically generates new responses for calibration prompts, rather than using static pre-existing answers.
- **Normalization**: Formatting calibration samples using the target model's official chat template to align activations with the instruction-tuned distribution.

4.2 Calibration Regeneration

We believe that models drawn from the same developmental lineage and trained under similar paradigms tend to exhibit consistent internal activations. This motivates our core hypothesis: calibration data regenerated by a more capable, in-family model are more effective for PTQ than real-world data alone. An in-family "Elder Sibling" model, sharing architecture and training paradigms with the target, implicitly encodes the target's activation dynamics (see Appendix Figure 4). Consequently, data it generates are not only semantically rich but also tailored to drive the target model's inference in a representative way, including surface-level outliers. This alignment yields more accurate quantization parameters and mitigates accuracy loss.

To exploit this prior, we introduce **In-family Synthesis with CoT**. For each query in the original seed calibration set, we prompt the larger, in-family "Elder Sibling" model to generate a new, detailed response. Crucially, we instruct the model to produce not just a final answer but also its intermediate reasoning process (e.g., via CoT prompting). This

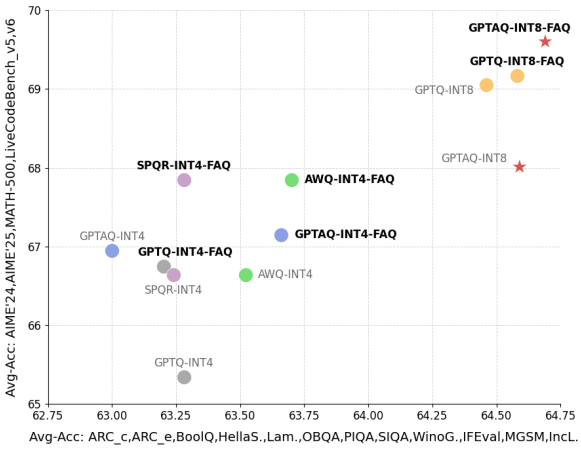


Figure 3: Overall performance improvement of FAQ across multiple quantization methods and benchmark suites on the Qwen3-8B model. X: average accuracy on general tasks; Y: performance on math/coding tasks. Each point is a quantization method (color = base algorithm, e.g., green for AWQ-INT4; gray vs bold label = baseline vs FAQ). The consistent up-right shift shows FAQ is a plug-and-play enhancement.

yields data with greater complexity and semantic diversity, designed to engage a broader set of pathways in the target model.

4.3 Calibration Normalization

The regeneration process produces a large number of synthetic, activation-distribution-aligned calibration samples. However, normalization is still necessary in practical applications to avoid the negative impact of unreasonable and unstable data on PTQ.

First, we designed **Quality-driven Selection.** We generate three candidate responses per query and use a powerful external LLM (e.g., Qwen2.5-72B-Instruct (Yang et al., 2024)) as a judge to select the highest-scoring one, filtering out nonsensical generations. Secondly, **Template-based Data Assembly.** The selected response is combined with the original query and formatted using the target model's official chat template. This ensures the final sample's format perfectly aligns with the model's expected input, enhancing distributional consistency.

Together, regeneration and normalization convert a small seed set into a diverse, distributionally aligned calibration corpus that is structurally compatible with the target model, thereby enhancing PTQ performance.

5 Experiments

In this section, we conduct a comprehensive set of experiments to validate FAQ's effectiveness and versatility as a plug-and-play enhancement for PTQ methods. To provide a holistic overview of our findings, we begin with Figure 3, which encapsulates the main results on the Qwen3-8B model. As the plot demonstrates, applying our FAQ framework shifts the performance of baseline methods towards the top-right corner, signifying improved accuracy of the model. This overarching improvement provides the context for the detailed numerical results and analyses presented in the subsequent sections.

Bits	Method	Qwen3-8B		
		Wiki2(↓)	C4(↓)	Lambada(↓)
BF16	-	12.20	36.37	6.19
INT8	GPTQ	12.20	36.36	6.25
	+FAQ	**12.19**	**36.33**	**6.23**
	GPTAQ	12.22	**36.37**	6.21
	+FAQ	**12.22**	36.40	**6.16**
INT4	GPTQ	13.10	38.97	7.55
	+FAQ	**12.95**	**38.62**	**7.26**
	AWQ	12.80	38.72	**6.90**
	+FAQ	**12.73**	**38.60**	7.16
	SPQR	16.40	56.70	**7.37**
	+FAQ	**14.95**	**46.27**	8.03
	GPTAQ	13.01	39.17	6.88
	+FAQ	**12.99**	**39.01**	**6.62**
AVG	Quant	13.29	41.05	**6.86**
	+FAQ	**13.01**	**39.21**	6.91

Table 1: Perplexity comparison of different quantization methods on Qwen3-8B models. The lower is the better.

5.1 Experimental Setup

Models and Methods. We evaluate our framework on a diverse suite of recent open-source models from the Qwen3 family, including the dense Qwen3-8B, a reasoning-distilled variant, and the larger MoE-based Qwen3-30B-A3B. This allows us to assess performance across standard, specialized, and sparse architectures. We apply FAQ as a plug-and-play enhancement to four PTQ methods: GPTQ, AWQ, SPQR, and GPTAQ, primarily in INT4 and INT8 settings.

Evaluation. Our comprehensive evaluation covers three key areas:

1. Language Modeling, measured by perplexity on Wikitext2 (Merity et al., 2017), C4 (Raffel et al., 2020), and LAMBADA (Radford et al., 2019) benchmarks;

Bits	Method	ARC_c	ARC_e	BoolQ	HellaS.	Lam.	OBQA	PIQA	SIQA	WinoG.	IFEval	MGSM	IncL.	Avg
BF16	-	55.2	83.6	86.7	57.1	61.6	31.4	76.6	41.8	68.0	84.8	54.6	75.4	64.7
INT8	GPTQ	**55.1**	**83.5**	86.6	**57.2**	61.4	**31.4**	76.8	41.5	67.7	83.0	54.0	75.4	64.5
	+FAQ	55.0	83.4	**86.6**	57.1	**61.5**	31.1	**76.9**	**41.6**	**68.1**	**83.2**	**54.9**	**75.8**	**64.6**
	GPTAQ	55.0	83.6	**86.5**	**57.1**	61.4	**31.6**	76.6	41.4	**68.3**	83.3	54.5	**75.8**	64.6
	+FAQ	**55.5**	**83.7**	86.4	**57.1**	**61.7**	31.5	**76.7**	**41.6**	68.0	**83.8**	**54.7**	75.6	**64.7**
INT4	GPTQ	**52.7**	80.9	**86.3**	55.9	59.0	29.8	76.3	41.1	**68.5**	81.3	**56.4**	**71.2**	**63.3**
	+FAQ	51.8	**81.6**	85.8	**55.9**	**59.6**	**31.2**	**76.8**	**42.1**	68.2	**82.1**	53.7	69.7	63.2
	AWQ	52.8	81.4	**86.7**	56.0	60.1	30.2	**76.2**	**41.0**	**67.5**	82.1	54.2	**73.9**	63.5
	+FAQ	**53.1**	**81.6**	86.4	**56.0**	**60.6**	**30.8**	76.1	40.5	67.0	**83.6**	**56.6**	71.8	**63.7**
	SPQR	**54.5**	**83.0**	85.9	**56.4**	59.3	29.8	76.0	**40.8**	66.8	82.8	**53.3**	70.3	63.2
	+FAQ	53.5	80.9	**86.1**	56.0	**59.8**	**30.4**	**77.3**	40.5	**68.4**	**83.6**	51.8	**71.1**	**63.3**
	GPTAQ	53.0	81.5	85.3	55.3	**60.1**	**31.6**	75.7	41.0	66.9	82.3	51.4	71.9	63.0
	+FAQ	**53.8**	**82.0**	**86.4**	**55.8**	60.0	29.6	**75.9**	**41.0**	**68.6**	**83.1**	**53.8**	**74.0**	**63.7**
AVG	Quant	**53.9**	**82.3**	86.2	56.3	60.2	30.7	76.2	41.1	67.6	82.5	54.0	**73.1**	63.7
	+FAQ	53.8	82.2	**86.3**	**56.3**	**60.5**	**30.8**	**76.6**	**41.2**	**68.1**	**83.2**	**54.2**	73.0	**63.9**

Table 2: Performance on 12 general downstream tasks with different quantization methods on Qwen3-8B models. Higher is better. The positive trend in the average score suggests that the benefits of FAQ are systematic.

2. General Reasoning and Multilingual Capabilities, assessed via the average accuracy on a broad suite of 12 downstream tasks:ARC-c and ARC-e (Clark et al., 2018), BoolQ (Clark et al., 2019), Hellaswag (Zellers et al., 2019), LAMBADA, OpenBookQA (Mihaylov et al., 2018), PIQA (Bisk et al., 2020), SIQA (Sap et al., 2019), Winogrande (Sakaguchi et al., 2021), IFEval (Zhou et al., 2023), MGSM (Shi et al., 2023) and INCLUDE (Romanou et al., 2025);

3. Specialized Domain Capabilities, tested on challenging math and code generation benchmarks: AIME'24 and AIME'25 (AIME, 2025), MATH-500 (Lightman et al., 2024), and LiveCodeBench (Jain et al., 2025).

All downstream evaluations are conducted in a zero-shot setting. To ensure robustness, all reported scores are averaged over multiple runs. A detailed description of models, PTQ configurations, benchmark lists, and our reliability assurance protocol can be found in Appendix A.

5.2 Main Results

5.2.1 Language Modeling Performance.

We first evaluate the impact of FAQ on the fundamental language modeling capabilities of the quantized model, measured by Perplexity (PPL) on the Wikitext2, C4, and LAMBADA datasets. Lower PPL indicates better performance. The detailed results for the Qwen3-8B model are presented in Table 1. The results demonstrate that FAQ enhances the language modeling performance of all evaluated PTQ methods across both INT8 and INT4 quantization in most cases. The average (AVG) PPL across all three benchmarks is consistently reduced with FAQ. For instance, in the more challenging INT4 setting, FAQ brings a substantial PPL reduction for SPQR on the C4 dataset, decreasing it from 56.70 to 46.27. Similarly, for GPTQ-INT4 and GPTAQ-INT4, FAQ improves performance across all three datasets, highlighting its robust and broad effectiveness. This consistent improvement can be attributed to the higher-quality calibration data generated by FAQ. By producing a calibration set that better reflects the model's typical activation distributions and tames outlier features, FAQ enables the quantizer to learn more accurate scaling factors. This leads to lower quantization error and, consequently, a better preservation of the model's nuanced understanding of language, as evidenced by the lower perplexity scores.

5.2.2 General Reasoning and Multilingual Capabilities.

To assess whether improvements in language modeling translate to broader cognitive abilities, we evaluate the models on a diverse suite of 12 downstream tasks. As detailed in Table 2, the results demonstrate the wide-ranging benefits of FAQ. On average, applying FAQ leads to performance gains, reducing the accuracy loss compared to the full-

precision baseline by 28.5%.

To contextualize this improvement, we introduce the **Relative Error Reduction (RER)** metric:

$$\text{RER} = \frac{\text{Score}_{\text{FAQ}} - \text{Score}_{\text{Baseline}}}{\text{Score}_{\text{BF16}} - \text{Score}_{\text{Baseline}}} \times 100\%, \quad (6)$$

which measures the proportion of quantization-induced accuracy loss that FAQ recovers. On the general reasoning suite, FAQ achieves an RER of approximately 28.5%, recovering nearly one-third of the INT4 degradation.

This positive trend is particularly notable on reasoning-intensive benchmarks like IFEval (instruction-following) and MGSM (multilingual math), where FAQ helps recover significant performance otherwise lost to quantization. This suggests that FAQ's enhanced calibration preserves not just language modeling fidelity but also the critical model parameters responsible for complex reasoning, enabling the quantized model to better retain its general-purpose problem-solving skills.

We note a marginal 0.1-point average accuracy dip for the GPTQ-INT4 configuration, which we trace to a significant performance drop solely on the MGSM benchmark. We hypothesize this is a corner-case interaction where our regeneration process, in an otherwise beneficial act of smoothing distributions for the highly aggressive GPTQ-INT4 quantizer, may have inadvertently altered outlier features that were coincidentally crucial for MGSM's specific numerical reasoning patterns. This finding highlights a complex trade-off in extreme quantization scenarios, but the strong positive trend across all other methods and benchmarks confirms the substantial net benefit of FAQ.

5.2.3 Specialized Domain Capabilities: Math and Code.

Finally, to rigorously probe the limits of quantized models, we evaluate them on highly challenging math (AIME, MATH-500) and code generation (LiveCodeBench) benchmarks, which are notoriously sensitive to precision loss. The results in Table 3 demonstrate the remarkable resilience FAQ provides. Quantitative results show FAQ improves accuracy from 67.0 to 68.0 while reducing quantization-induced accuracy loss by 22%, conclusively demonstrating that our data regeneration strategy effectively preserves models' complex reasoning abilities.

The benefits are particularly evident in the most aggressive INT4 setting. For instance, FAQ boosts the average accuracy of AWQ by 1.3 points and SPQR by a notable 1.4 points, with consistent gains observed across individual benchmarks like AIME and LiveCodeBench. This strong performance provides compelling evidence for our core hypothesis: complex reasoning relies on high-magnitude activations that represent critical operations. By generating data that tames the distribution of these crucial features, FAQ prevents their distortion during quantization, ensuring the essential building blocks of the model's reasoning capabilities are kept intact, even under extreme compression.

Bits	Method	AIME		Math	L.C.B.		Avg
		24(↑)	**25**(↑)	**500**(↑)	**v5**(↑)	**v6**(↑)	
BF16	-	83.3	76.7	95.2	58.1	52.6	73.2
INT8	GPTQ	76.3	**68.3**	94.6	57.2	48.9	69.1
	+FAQ	**79.2**	62.9	**95.1**	**58.7**	**50.0**	**69.2**
	GPTAQ	74.6	68.8	**94.7**	53.3	48.7	68.0
	+FAQ	**78.3**	**69.2**	94.2	**56.3**	**50.0**	**69.6**
INT4	GPTQ	**74.2**	59.2	95.0	**52.1**	46.3	65.3
	+FAQ	73.3	**66.7**	**95.1**	51.5	**47.2**	**66.8**
	AWQ	72.5	63.3	94.9	53.9	48.6	66.6
	+FAQ	**75.4**	**65.0**	**94.9**	**55.1**	**48.9**	**67.9**
	SPQR	74.6	65.6	94.2	50.6	45.4	66.1
	+FAQ	**76.7**	**65.6**	**94.4**	**54.2**	**46.4**	**67.5**
	GPTAQ	**78.3**	62.5	94.8	52.7	46.4	67.0
	+FAQ	75.0	**64.4**	**94.8**	**53.3**	**48.2**	**67.2**
AVG	Quant	75.1	64.6	94.7	53.3	47.4	67.0
	+FAQ	**76.3**	**65.6**	**94.8**	**54.8**	**48.4**	**68.0**

Table 3: Accuracy on specialized math and code benchmarks for the Qwen3-8B model. FAQ's consistent performance improvement, especially in the challenging INT4 setting, demonstrates its ability to preserve critical reasoning capabilities under extreme compression. Higher is better.

5.3 Contamination Analysis

A critical question in calibration data design is whether the observed improvements stem from genuine distribution alignment or inadvertent data contamination. To rigorously address this, we conduct an **Oracle Calibration** experiment, using the AIME 2024 test questions themselves as calibration data. If the gains from FAQ were attributable to data leakage, this oracle setup should yield superior results.

As shown in Table 4, the Oracle Calibration setup achieves an average accuracy of 69.62, no-

Method	AIME-24 (↑)
Oracle Calibration	69.62
GPTQ-INT4 + FAQ	**73.33**

Table 4: Contamination analysis via Oracle Calibration on Qwen3-8B INT4. Higher is better.

tably lower than FAQ's 73.33. This result decisively demonstrates that (i) directly using test questions does not confer an advantage, as they fail to provide the balanced activation coverage needed for robust quantization, and (ii) FAQ's improvements are genuinely attributable to its family-aware distribution alignment strategy, not to any form of data contamination.

5.4 Component Ablation Study

To understand the contribution of each component in the FAQ pipeline, we conduct a comprehensive ablation study on Qwen3-8B with GPTQ-INT4. We systematically remove individual components while keeping the rest of the pipeline intact.

Method	AIME		Math	LCB	Avg
	24	25	500	v5	
Full	**73.33**	**66.67**	**95.10**	**51.50**	**71.65**
w/o Judge	69.38	62.30	94.09	51.50	69.31 (↓2.34)
w/o CoT	73.33	63.34	94.30	50.98	70.49 (↓1.16)
w/o Chat-T	71.46	60.00	93.95	50.45	70.36 (↓1.29)

Table 5: Component ablation study of FAQ on Qwen3-8B GPTQ-INT4. Each variant removes one component while keeping others fixed. Higher is better.

The ablation results Table 5 reveal several key insights. First, removing the judge model (using only 1-of-1 generation without selection) causes the largest average accuracy drop of 2.34 points, confirming that quality filtering is the most critical component of FAQ. Second, removing chain-of-thought reasoning during regeneration reduces performance by 1.16 points, indicating that CoT provides useful intermediate activation patterns for quantization calibration. Third, omitting the chat template normalization leads to a 1.29-point drop, demonstrating that input format alignment with the model's instruction-tuned distribution is essential for optimal calibration.

5.5 Hypothesis Validation and Generalization

Having established the significant and consistent benefits of FAQ on the Qwen3-8B model, we now turn to two critical questions to further understand its underlying principles and scope of applicability. First, we conduct an ablation study to validate our core "Family-Aware" hypothesis. Second, we test the generalization of FAQ to a larger-scale, more complex MoE model to assess its scalability.

Bits	Method	DeepSeek-R1-0528-Qwen3-8B				
		Wiki2	C4	Lam.	GRMC	SDC
BF16	-	13.23	40.83	27.32	60.64	71.57
INT8	GPTQ+ds	13.24	**40.85**	27.42	60.74	70.44
	GPTQ+qw	**13.24**	40.86	**27.41**	**60.77**	**70.89**
	GPTAQ+ds	**13.20**	40.73	27.32	60.42	69.80
	GPTAQ+qw	13.21	**40.72**	**27.12**	**60.51**	**69.93**
INT4	GPTQ+ds	14.30	44.27	33.95	58.26	67.14
	GPTQ+qw	**14.22**	**44.24**	**31.90**	**58.27**	**68.26**
	AWQ+ds	13.93	43.29	33.65	59.19	66.83
	AWQ+qw	**13.87**	**43.00**	**30.11**	**59.27**	**68.20**
	SPQR+ds	14.36	**44.18**	34.61	58.00	65.16
	SPQR+qw	**14.35**	44.58	**29.77**	**58.28**	**67.11**
	GPTAQ+ds	14.37	**44.12**	33.66	58.98	67.69
	GPTAQ+qw	**14.31**	44.18	**29.89**	**59.65**	**68.03**
AVG	Quant+ds	13.90	**42.91**	31.77	59.27	67.84
	Quant+qw	**13.87**	42.93	**29.37**	**59.46**	**68.74**

Table 6: Performance comparison of Family-Sourced (+qw) vs. Knowledge-Sourced (+ds) calibration data on the distilled DeepSeek-R1-0528-Qwen3-8B model. Lower perplexity is better; higher accuracy is better (GRMC/SDC). +qw consistently outperforms +ds, supporting the importance of within-family alignment for FAQ.

5.5.1 Validating the "Family-Aware" Hypothesis.

A Unique Testbed: The Distilled Model. Our core hypothesis posits that the efficacy of FAQ stems from a deep "family" connection, which we define not by macro-architectural identity (e.g., MoE vs. Dense) but by a shared developmental lineage. This includes overlapping training corpora, consistent tokenization schemes, and a common design philosophy for fundamental building blocks. To rigorously test this, we leverage a unique model: DeepSeek-R1-0528-Qwen3-8B. This model was created by distilling knowledge from a "knowledge teacher," DeepSeek-R1 (an MoE model), into a

"student model," Qwen3-8B (a Dense model). This setup creates a fascinating and powerful adversarial test. We compare two data generation strategies, both using large MoE models as generators:

1. Teacher-Sourced Calibration (+ds): Generating data using the knowledge teacher, DeepSeek-R1.

2. Family-Sourced Calibration (+qw): Generating data using the student's family model, Qwen3-235B-A22B.

Results and Analysis. The results in Table 6 provide a decisive answer to the central question of whether knowledge origin (DeepSeek) or developmental lineage (Qwen) is more effective. The Family-Sourced (+qw) calibration decisively outperforms the Teacher-Sourced (+ds) data, improving the average score on specialized tasks (68.74 vs. 67.84) and reducing perplexity on Lambada (29.37 vs. 31.77).

This strong head-to-head comparison validates our core hypothesis: a shared "developmental lineage" is more critical than knowledge origin. The superior performance is not attributable to a naive architectural match—as both generators are MoE while the student is Dense—but rather stems from a form of "distributional homology". This homology enables the generator to craft inputs that elicit more natural, quantization-friendly activation patterns. Due to space constraints, we present the averaged results here; a detailed, per-benchmark comparison can be found in Appendix.

5.5.2 Generalization to MoE Architectures.

To demonstrate FAQ's scalability and versatility, we apply it to the Qwen3-30B-A3B, a larger MoE model whose sparse, gated activations present a rigorous test. The results in Table 7 robustly mirror our earlier success. On average, FAQ boosts the specialized task score from 67.84 to 68.74, confirming its effectiveness on complex architectures.

The performance lift is particularly pronounced in the demanding INT4 setting, where FAQ boosts SPQR's accuracy by nearly two points (65.16 to 67.11) and GPTQ's from 67.14 to 68.26. This is also reflected in a substantial perplexity reduction on Lambada. This successful application to a 30B MoE model validates that FAQ is a robust and generalizable framework, whose benefits are not confined to model size or density but extend effectively to sparse activation environments, making it a broadly applicable tool for modern LLMs.

Bits	Method	Qwen3-30B-A3B				
		Wiki2	C4	Lam.	GRMC	SDC
BF16	-	10.89	30.13	5.44	66.85	74.37
INT8	GPTQ	11.02	30.84	5.58	66.55	71.50
	+FAQ	**11.02**	**30.84**	**5.56**	**66.81**	**72.29**
INT4	GPTQ	11.36	31.87	6.21	65.37	71.53
	+FAQ	**11.32**	**31.60**	**6.14**	**65.53**	**71.78**
	SPQR	13.08	40.49	6.55	65.10	70.69
	+FAQ	**11.83**	**34.81**	**6.01**	**65.36**	**71.66**
AVG	Quant	11.82	34.40	6.11	65.67	71.24
	+FAQ	**11.39**	**32.42**	**5.90**	**65.90**	**71.91**

Table 7: Generalization of FAQ to the larger-scale, MoE-based Qwen3-30B-A3B model. We compare baseline PTQ methods and their FAQ-enhanced variants; FAQ consistently improves perplexity and GRMC/SDC accuracy, demonstrating scalability to larger, more complex architectures.

6 Conclusions

We introduce FAQ, a **Q**uantization framework that optimizes calibration data by exploiting **F**amily **A**wareness. FAQ addresses a fundamental source of quantization error—outlier-prone activations—by leveraging an in-family, larger "Elder Sibling" model to regenerate a high-quality, quantization-friendly calibration set. Comprehensive experiments demonstrate that FAQ acts as a universal, plug-and-play performance booster across a broad spectrum of SOTA PTQ methods, model sizes, and architectures, including MoE models. Crucially, our ablation study provides direct evidence that FAQ's effectiveness stems from a shared developmental lineage between models, a factor more critical than the origin of knowledge or macro-architectural similarity. Collectively, this work inaugurates a data-centric paradigm for PTQ and highlights the substantial potential of leveraging model family priors to yield more robust and efficient LLMs.

Limitations

The effectiveness of FAQ can be influenced by specific choices in the regeneration process, such as prompting strategies and generation hyperparameters (e.g., temperature, candidate selection). While our default settings demonstrate broad robustness, extreme quantization regimes (e.g., INT4) inherently introduce volatility. Consequently, minor performance fluctuations may occur in isolated edge

cases—such as the marginal degradation observed on MGSM—suggesting a valuable avenue for future research into adaptive calibration strategies to further enhance stability.

References

AIME. 2025. Aime problems and solutions. https://artofproblemsolving.com/wiki/index.php/AIMEProblemsandSolutions.

Yonatan Bisk, Rowan Zellers, Ronan Le Bras, Jianfeng Gao, and Yejin Choi. 2020. PIQA: reasoning about physical commonsense in natural language. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*, pages 7432–7439. AAAI Press.

Yaohui Cai, Zhewei Yao, Zhen Dong, Amir Gholami, Michael W. Mahoney, and Kurt Keutzer. 2020. Zeroq: A novel zero shot quantization framework. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*, pages 13166–13175. Computer Vision Foundation / IEEE.

Jerry Chee, Yaohui Cai, Volodymyr Kuleshov, and Christopher De Sa. 2023. Quip: 2-bit quantization of large language models with guarantees. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.

Yoni Choukroun, Eli Kravchik, Fan Yang, and Pavel Kisilev. 2019. Low-bit quantization of neural networks for efficient inference. In *2019 IEEE/CVF International Conference on Computer Vision Workshops, ICCV Workshops 2019, Seoul, Korea (South), October 27-28, 2019*, pages 3009–3018. IEEE.

Christopher Clark, Kenton Lee, Ming-Wei Chang, Tom Kwiatkowski, Michael Collins, and Kristina Toutanova. 2019. 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, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 2924–2936. Association for Computational Linguistics.

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. 2018. Think you have solved question answering? try arc, the ai2 reasoning challenge. *arXiv preprint arXiv:1803.05457*.

Tim Dettmers, Ruslan Svirschevski, Vage Egiazarian, Denis Kuznedelev, Elias Frantar, Saleh Ashkboos, Alexander Borzunov, Torsten Hoefler, and Dan Alistarh. 2024. Spqr: A sparse-quantized representation for near-lossless LLM weight compression. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net.

Vage Egiazarian, Andrei Panferov, Denis Kuznedelev, Elias Frantar, Artem Babenko, and Dan Alistarh. 2024. Extreme compression of large language models via additive quantization. In *Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024*. OpenReview.net.

Steven K. Esser, Jeffrey L. McKinstry, Deepika Bablani, Rathinakumar Appuswamy, and Dharmendra S. Modha. 2020. Learned step size quantization. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Elias Frantar, Saleh Ashkboos, Torsten Hoefler, and Dan Alistarh. 2023. OPTQ: accurate quantization for generative pre-trained transformers. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net.

Suriya Gunasekar, Yi Zhang, Jyoti Aneja, Caio César Teodoro Mendes, Allie Del Giorno, Sivakanth Gopi, Mojan Javaheripi, Piero Kauffmann, Gustavo de Rosa, Olli Saarikivi, Adil Salim, Shital Shah, Harkirat Singh Behl, Xin Wang, Sébastien Bubeck, Ronen Eldan, Adam Tauman Kalai, Yin Tat Lee, and Yuanzhi Li. 2023. Textbooks are all you need. *CoRR*, abs/2306.11644.

Wei Huang, Haotong Qin, Yangdong Liu, Yawei Li, Qinshuo Liu, Xianglong Liu, Luca Benini, Michele Magno, Shiming Zhang, and Xiaojuan Qi. 2025. Slim-llm: Salience-driven mixed-precision quantization for large language models. In *Forty-second International Conference on Machine Learning, ICML 2025, Vancouver, BC, Canada, July 13-19, 2025*. OpenReview.net.

Itay Hubara, Yury Nahshan, Yair Hanani, Ron Banner, and Daniel Soudry. 2020. Improving post training neural quantization: Layer-wise calibration and integer programming. *arXiv preprint arXiv:2006.10518*.

Benoit Jacob, Skirmantas Kligys, Bo Chen, Menglong Zhu, Matthew Tang, Andrew G. Howard, Hartwig Adam, and Dmitry Kalenichenko. 2018. Quantization and training of neural networks for efficient integer-arithmetic-only inference. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*, pages 2704–2713. Computer Vision Foundation / IEEE Computer Society.

Naman Jain, King Han, Alex Gu, Wen-Ding Li, Fanjia Yan, Tianjun Zhang, Sida Wang, Armando Solar-Lezama, Koushik Sen, and Ion Stoica. 2025. Livecodebench: Holistic and contamination free evalua-

tion of large language models for code. In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net.

Guochao Jiang, Guofeng Quan, Zepeng Ding, Ziqin Luo, Dixuan Wang, and Zheng Hu. 2025. Flashthink: An early exit method for efficient reasoning. *arXiv preprint arXiv:2505.13949*.

Sehoon Kim, Coleman Hooper, Amir Gholami, Zhen Dong, Xiuyu Li, Sheng Shen, Michael W. Mahoney, and Kurt Keutzer. 2024. Squeezellm: Dense-and-sparse quantization. In *Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024*. OpenReview.net.

Changhun Lee, Jungyu Jin, Taesu Kim, Hyungjun Kim, and Eunhyeok Park. 2024. OWQ: outlier-aware weight quantization for efficient fine-tuning and inference of large language models. In *Thirty-Eighth AAAI Conference on Artificial Intelligence, AAAI 2024, Thirty-Sixth Conference on Innovative Applications of Artificial Intelligence, IAAI 2024, Fourteenth Symposium on Educational Advances in Artificial Intelligence, EAAI 2014, February 20-27, 2024, Vancouver, Canada*, pages 13355–13364. AAAI Press.

Jung Hyun Lee, Jeonghoon Kim, Se Jung Kwon, and Dongsoo Lee. 2023. Flexround: Learnable rounding based on element-wise division for post-training quantization. In *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, volume 202 of *Proceedings of Machine Learning Research*, pages 18913–18939. PMLR.

Yuhang Li, Ruihao Gong, Xu Tan, Yang Yang, Peng Hu, Qi Zhang, Fengwei Yu, Wei Wang, and Shi Gu. 2021a. BRECQ: pushing the limit of post-training quantization by block reconstruction. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net.

Yuhang Li, Ruihao Gong, Xu Tan, Yang Yang, Peng Hu, Qi Zhang, Fengwei Yu, Wei Wang, and Shi Gu. 2021b. BRECQ: pushing the limit of post-training quantization by block reconstruction. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net.

Yuhang Li, Ruokai Yin, Donghyun Lee, Shiting Xiao, and Priyadarshini Panda. 2025. GPTAQ: efficient finetuning-free quantization for asymmetric calibration. In *Forty-second International Conference on Machine Learning, ICML 2025, Vancouver, BC, Canada, July 13-19, 2025*. OpenReview.net.

Zhongzhi Li, Xuansheng Wu, Yijiang Li, Lijie Hu, and Ninghao Liu. 2026. Less is enough: Synthesizing diverse data in feature space of llms. *CoRR*, abs/2602.10388.

Hunter Lightman, Vineet Kosaraju, Yuri Burda, Harrison Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. 2024. Let's verify step by step. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net.

Ji Lin, Jiaming Tang, Haotian Tang, Shang Yang, Guangxuan Xiao, and Song Han. 2024. AWQ: activation-aware weight quantization for on-device LLM compression and acceleration. *GetMobile Mob. Comput. Commun.*, 28(4):12–17.

Stephen Merity, Caiming Xiong, James Bradbury, and Richard Socher. 2017. Pointer sentinel mixture models. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net.

Todor Mihaylov, Peter Clark, Tushar Khot, and Ashish Sabharwal. 2018. 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, Brussels, Belgium, October 31 - November 4, 2018*, pages 2381–2391. Association for Computational Linguistics.

Lin Ou, Jinpeng Xia, Yuewei Zhang, Chuzhan Hao, and Hao Henry Wang. 2024. Adaptive quantization error reconstruction for llms with mixed precision. In *First Conference on Language Modeling*.

Guofeng Quan, Wenfeng Feng, Chuzhan Hao, Guochao Jiang, Yuewei Zhang, and Hao Henry Wang. 2025. RASD: retrieval-augmented speculative decoding. In *Findings of the Association for Computational Linguistics, ACL 2025, Vienna, Austria, July 27 - August 1, 2025*, pages 6167–6177. Association for Computational Linguistics.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, and 1 others. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

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

Angelika Romanou, Negar Foroutan, Anna Sotnikova, Zeming Chen, Sree Harsha Nelaturu, Shivalika Singh, Rishabh Maheshwary, Micol Altomare, Mohamed A. Haggag, Imanol Schlag, Marzieh Fadaee, Sara Hooker, Antoine Bosselut, Snegha A, Alfonso Amayuelas, Azril Hafizi Amirudin, Viraat Aryabumi, Danylo Boiko, Michael Chang, and 40 others. 2025. INCLUDE: evaluating multilingual language understanding with regional knowledge. In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net.

Keisuke Sakaguchi, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. 2021. Winogrande: an adversarial winograd schema challenge at scale. *Commun. ACM*, 64(9):99–106.

Maarten Sap, Hannah Rashkin, Derek Chen, Ronan LeBras, and Yejin Choi. 2019. Socialiqa: Commonsense reasoning about social interactions. *arXiv preprint arXiv:1904.09728*.

Wenqi Shao, Mengzhao Chen, Zhaoyang Zhang, Peng Xu, Lirui Zhao, Zhiqian Li, Kaipeng Zhang, Peng Gao, Yu Qiao, and Ping Luo. 2024. Omniquant: Omnidirectionally calibrated quantization for large language models. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net.

Freda Shi, Mirac Suzgun, Markus Freitag, Xuezhi Wang, Suraj Srivats, Soroush Vosoughi, Hyung Won Chung, Yi Tay, Sebastian Ruder, Denny Zhou, Dipanjan Das, and Jason Wei. 2023. Language models are multilingual chain-of-thought reasoners. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net.

Boris van Breugel, Yelysei Bondarenko, Paul Whatmough, and Markus Nagel. 2025. Fptquant: Function-preserving transforms for llm quantization. *arXiv preprint arXiv:2506.04985*.

Guangxuan Xiao, Ji Lin, Mickaël Seznec, Hao Wu, Julien Demouth, and Song Han. 2023. Smoothquant: Accurate and efficient post-training quantization for large language models. In *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, volume 202 of *Proceedings of Machine Learning Research*, pages 38087–38099. PMLR.

Can Xu, Qingfeng Sun, Kai Zheng, Xiubo Geng, Pu Zhao, Jiazhan Feng, Chongyang Tao, Qingwei Lin, and Daxin Jiang. 2024. Wizardlm: Empowering large pre-trained language models to follow complex instructions. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, and 1 others. 2024. Qwen2. 5 technical report. *arXiv preprint arXiv:2412.15115*.

Ping Yu, Jack Lanchantin, Tianlu Wang, Weizhe Yuan, Olga Golovneva, Ilia Kulikov, Sainbayar Sukhbaatar, Jason Weston, and Jing Xu. 2025. Cot-self-instruct: Building high-quality synthetic prompts for reasoning and non-reasoning tasks. *CoRR*, abs/2507.23751.

Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. 2019. Hellaswag: Can a machine really finish your sentence? In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 4791–4800. Association for Computational Linguistics.

Jeffrey Zhou, Tianjian Lu, Swaroop Mishra, Siddhartha Brahma, Sujoy Basu, Yi Luan, Denny Zhou, and Le Hou. 2023. Instruction-following evaluation for large language models. *arXiv preprint arXiv:2311.07911*.

Appendix

A Detailed Experimental Setup

A.1 Models and Quantization Methods.

Our investigation is conducted on a diverse set of recent open-source models with distinct architectures, selected based on the following rationale:

- **Qwen3 Ecosystem**: Provides the most comprehensive range of publicly available model sizes (0.5B to 235B) sharing a common tokenizer and training pipeline.
- **Qwen3-8B / 30B**: Represent the critical "sweet spot" for open-source deployment. Qwen3-30B-A3B is a Mixture-of-Experts (MoE) model, enabling evaluation across dense and sparse architectures.
- **DeepSeek-R1**: A reasoning-distilled variant (DeepSeek-R1-0528-Qwen3-8B) created by distilling from a larger MoE teacher into a dense student, serving as a unique adversarial testbed for our family hypothesis.
- **Cross-Family Validation**: Llama3.1-8B experiments demonstrate FAQ's generalizability beyond the Qwen ecosystem.

We benchmark four post-training quantization (PTQ) algorithms with our proposed FAQ: GPTQ, AWQ, SPQR, and GPTAQ. For GPTQ and GPTAQ, we explore both INT4 and INT8 precision levels, while AWQ and SPQR are evaluated in their standard INT4 configuration. Due to specific library or architectural constraints, particularly for the MoE model, Qwen3-30B-A3B is benchmarked using only the GPTQ and SPQR methods. For all quantization procedures, we consistently use a group_size of 128 and a calibration set composed of 256 samples, as these are common and effective settings in the PTQ literature.

Method	ARC_c	ARC_e	BoolQ	HellaS.	Lam.	OBQA	PIQA	SIQA	WinoG.	IFEval	MGSM	IncL.	Avg
BF16	52.05	80.3	86.39	58.41	36.89	31	76.06	43.09	62.12	73.01	58.1	70.28	60.64
INT8													
GPTQ+ds	**52.31**	79.97	**86.38**	58.32	36.72	**31.1**	76.25	**43.09**	**62.35**	74.45	**58.25**	69.73	60.74
GPTQ+qw	52.14	**80.16**	86.33	**58.38**	**36.74**	30.9	**76.47**	42.84	62	**75.48**	58.05	**69.82**	**60.78**
GPTAQ+ds	52.3	80.09	**86.35**	**58.34**	36.67	**31.2**	76.2	**42.79**	61.64	75.33	53.93	**70.19**	60.42
GPTAQ+qw	**52.48**	**80.12**	86.33	58.24	**36.75**	30.6	**76.34**	42.48	**62.28**	**75.79**	**54.57**	70.18	**60.51**
INT4													
GPTQ+ds	50.64	**78.18**	**84.46**	**56.68**	35.36	**30**	74.65	41.05	**61.17**	70.15	**51.72**	**65.05**	58.26
GPTQ+qw	**51.07**	77.76	84.09	56.24	**36.7**	29.8	**74.95**	**41.74**	59.87	**70.89**	51.37	64.77	**58.27**
AWQ+ds	**50.56**	78.22	84.99	**57.05**	35.45	**31.7**	**76.34**	**42.35**	60.74	72.83	**51.92**	68.17	59.19
AWQ+qw	50.18	**78.43**	**85.14**	56.94	**36.68**	30.6	75.85	42.02	**61.84**	**73.39**	51.25	**68.9**	**59.27**
SPQR+ds	**51.84**	**78.26**	**84.65**	56.28	34.36	**30.2**	76.06	40.2	60.66	69.87	**49.78**	63.85	58.00
SPQR+qw	50.68	76.96	83.41	**56.93**	**36.69**	29.4	**76.12**	**41.33**	**60.69**	**73.57**	48.95	**64.59**	**58.28**
GPTAQ+ds	51.45	79.13	**85.38**	**56.28**	35.42	**30.7**	75.22	41.3	61.41	71.86	**53.52**	66.06	58.98
GPTAQ+qw	**52.09**	**79.15**	85.23	56.12	**37.35**	30.6	**76.55**	**42.15**	**62.47**	**73.78**	53.1	**67.25**	**59.65**
AVERAGE													
Quant+ds	**51.52**	**78.98**	**85.37**	**57.16**	35.66	**30.82**	75.79	41.80	61.33	72.42	**53.19**	67.18	59.27
Quant+qw	51.44	78.76	85.09	57.14	**36.82**	30.32	**76.05**	**42.09**	**61.53**	**73.82**	52.88	**67.59**	**59.46**

Table 8: Per-benchmark results on General tasks for DeepSeek-R1-0528-Qwen3-8B. Higher is better.

A.2 Evaluation Benchmarks and Metrics.

We perform a comprehensive, multi-faceted evaluation to assess model performance post-quantization. All evaluations on downstream tasks are conducted under a strict zero-shot setting.

- **Language Modeling**: We measure Perplexity (PPL) on Wikitext2, C4, and LAMBADA benchmarks.
- **General Reasoning and Multilingual Capabilities**: We evaluate performance on a broad suite of 12 downstream tasks: ARC-c and ARC-e, BoolQ, Hellaswag, LAMBADA, OpenBookQA, PIQA, SIQA, Winogrande, IFEval, MGSM and INCLUDE. For the benchmark IFEval, we report strict accuracy at the prompt level. For the multilingual math task MGSM, we use flexible-extract exact match. The INCLUDE benchmark, designed to evaluate regional knowledge, is tested specifically in its Chinese subset in our experiments.
- **Specialized Domain Capabilities**: To rigorously probe the limits of quantized models, we further tested them on highly challenging domain-specific benchmarks. For evaluating mathematical and logical reasoning skills,we employ high-level math benchmarks including professional exams AIME'24 and AIME'25, advanced mathematics MATH-500, and complex real-world coding challenges LiveCodeBench_v5, LiveCodeBench_v6.

For all accuracy-based evaluations on these tasks, we report the pass@1 metric, which means a single-attempt success rate.

B Detailed Experimental Results

This section provides the detailed, per-benchmark results for the experiments discussed in Subsections 5.5, Hypothesis Validation and Generalization, of the main paper. Due to space constraints, only the averaged results were presented in the main body. The following tables offer a granular view of the performance across all evaluated tasks, providing the full empirical evidence for our conclusions.

B.1 Detailed Results for the "Family-Aware" Hypothesis Validation

The following Table 8 and 9 detail the per-benchmark performance comparison of Family-Sourced (Qwen3-235B, denoted as '+qw') versus Knowledge-Sourced (DeepSeek-R1, denoted as '+ds') calibration data on the distilled DeepSeek-R1-0528-Qwen3-8B model.

Bits	Method	AIME'24	AIME'25	Math-500	LiveCodeBench_v5	LiveCodeBench_v6	Avg(↑)
BF16	-	86.67	65.63	93.6	61.08	50.86	71.57
INT8	GPTQ	82.92	66.15	93.2	**60.78**	49.14	70.44
	+FAQ	**84.17**	**68.65**	**93.5**	58.68	**49.43**	**70.89**
	GPTAQ	82.78	65.94	**92.93**	58.08	**49.29**	69.80
	+FAQ	**83.06**	**67.09**	92.67	**58.28**	48.57	**69.93**
INT4	GPTQ	81.67	62.61	**93.27**	53.59	44.57	67.14
	+FAQ	**83.06**	**63.20**	93.20	**56.59**	**45.29**	**68.27**
	AWQ	79.58	63.12	92.85	53.44	45.15	66.83
	+FAQ	**81.46**	**64.9**	**93.7**	**55.54**	**45.43**	**68.21**
	SPQR	79.79	**61.15**	91.9	52.69	40.29	65.16
	+FAQ	**81.25**	60.22	**93.5**	**56.89**	**43.72**	**67.12**
	GPTAQ	**80.42**	64.48	**93.2**	54.94	45.43	67.69
	+FAQ	80.83	**64.79**	93.1	**55.84**	**45.57**	**68.03**
AVG	Quant	81.19	63.91	92.89	55.59	45.65	67.85
	+FAQ	**82.31**	**64.81**	**93.28**	**56.97**	**46.34**	**68.74**

Table 9: Per-benchmark results on Specialized Domain tasks for DeepSeek-R1-0528-Qwen3-8B. Higher is better.

B.2 Detailed Results for Generalization to MoE Architectures

The following Table 10 and 11 present the granular, per-benchmark results for the Qwen3-30B-A3B MoE model, comparing baseline quantization methods against their FAQ-enhanced counterparts ('+FAQ').

C Detailed Implementation

Our evaluation pipeline is built on the lm-evaluation-harness and EvalScope toolkits to ensure standardized and reproducible assessment, as illustrated in code Listing 1, 2 and 3. High-performance inference is enabled by the SGLang and vLLM serving engines in code Listing 4 and Listing 5. To guarantee the statistical robustness and reliability of our findings, we adopt a stringent evaluation protocol. Each model-method configuration is evaluated 4 to 8 times on perplexity and general accuracy benchmarks, with the average score being reported. Recognizing the inherent high variance in complex reasoning tasks, we significantly increase the evaluation runs for the AIME datasets to a range of 32 to 64, reporting the averaged accuracy to provide a highly stable and credible performance measure. All experiments were performed on NVIDIA H20 GPUs.

C.1 Testing and Model Deployment Commands.

In the experiments of this paper, we used lm_eval[1] for evaluating the perplexity and general datasets, and evalscope[2] for evaluating the special dataset. The script commands used during the evaluations are shown in code listing 1, 2, and 3. When using lm_eval for evaluation, it is necessary to distinguish between the loglikelihood_rolling mode and the generate_until mode. Specifically, the generate_until mode is used for the IFEval and MGSM datasets, while the loglikelihood_rolling mode is used for the others.

To complete the dataset evaluation experiments faster, better, and more efficiently, we chose the mainstream inference service engines SGLang[3] and vLLM[4] for model deployment and inference. The corresponding model deployment command codes are shown in listing 4 and 5. It is important to note that due to the characteristics of the IFEval dataset evaluation, only vLLM can be used for deployment and API calls during testing.

Throughout the entire experiment, the versions of the relevant Python libraries used are as follows: torch 2.7.0, transformers 4.53.0, gptqmodel 4.0.0, sglang 0.4.9, vllm 0.9.1, lm_eval 0.4.9, and evalscope 2.0.0 .

[1] https://github.com/EleutherAI/lm-evaluation-harness
[2] https://github.com/modelscope/evalscope
[3] https://github.com/sgl-project/sglang
[4] https://github.com/vllm-project/vllm

Method	ARC_c	ARC_e	BoolQ	HellaS.	Lam.	OBQA	PIQA	SIQA	WinoG.	IFEval	MGSM	IncL.	Avg
BF16	52.39	79.50	88.62	59.61	63.89	34.40	79.38	43.14	70.09	85.58	61.57	84.04	66.85
INT8													
GPTQ	52.65	79.59	87.92	58.76	63.65	33.80	79.05	42.99	70.40	**84.66**	61.40	83.67	66.54
+FAQ	**52.74**	**79.65**	**88.15**	**58.78**	**63.65**	**34.90**	**79.55**	**43.76**	**70.80**	84.20	**61.40**	**84.13**	**66.81**
INT4													
GPTQ	**51.96**	77.61	87.77	58.16	**62.00**	**33.80**	78.84	**43.35**	68.82	82.26	**59.53**	80.28	65.36
+FAQ	50.51	**77.95**	**88.69**	**58.65**	61.81	33.40	**78.94**	42.89	**70.17**	**82.81**	58.17	**82.39**	**65.53**
SPQR	50.77	78.03	87.83	**58.02**	**61.42**	**33.60**	78.40	42.48	68.51	82.44	60.03	79.63	65.10
+FAQ	**51.54**	**78.70**	**87.83**	57.88	60.26	32.40	**78.67**	**42.84**	**68.90**	**83.55**	**60.43**	**81.28**	**65.36**
AVERAGE													
Quant	**51.79**	78.41	87.84	58.31	**62.36**	**33.73**	78.76	42.94	69.24	83.12	**60.32**	81.19	65.67
+FAQ	51.60	**78.77**	**88.22**	**58.44**	61.91	33.57	**79.05**	**43.16**	**69.96**	**83.52**	60.00	**82.60**	**65.90**

Table 10: Per-benchmark results on General tasks for Qwen3-30B-A3B. Higher is better.

Bits	Method	AIME'24	AIME'25	Math-500	LiveCodeBench_v5	LiveCodeBench_v6	Avg(↑)
BF16	-	81.67	71.88	95.40	63.47	59.43	74.37
INT8	GPTQ	79.72	68.06	94.85	59.28	**55.58**	71.50
	+FAQ	**80.56**	**70.11**	**94.85**	**60.48**	55.43	**72.28**
INT4	GPTQ	79.38	69.17	**94.75**	**60.63**	53.72	71.53
	+FAQ	**79.79**	**70.21**	94.68	59.88	**54.36**	**71.78**
	SPQR	**80.32**	68.23	94.70	57.78	52.43	70.69
	+FAQ	80.00	**68.44**	**95.50**	**61.08**	**53.29**	**71.66**
AVG	Quant	79.81	68.49	94.77	59.23	53.91	71.24
	+FAQ	**80.12**	**69.59**	**95.01**	**60.48**	**54.36**	**71.91**

Table 11: Per-benchmark results on Specialized Domain tasks for Qwen3-30B-A3B. Higher is better.

C.2 Computational Cost & Reproducibility

We provide a detailed breakdown of the computational overhead introduced by FAQ, as summarized in Table 12.

Several points are worth noting. First, the inference memory footprint during PTQ remains identical to the baseline (approximately 5.74 GB for Qwen3-8B GPTQ-INT4), as FAQ's generation and selection stages are performed offline. Second, the FAQ calibration dataset is reusable across multiple target models within the same family, amortizing the one-time generation cost. Finally, the seed dataset used for regeneration is NuminaMath-1.5, consisting of 256 math problems.

D Additional Experiments on Model Generalization and Data Generation Strategies

In this section, we present supplementary experiments designed to further explore the generalization capabilities and underlying mechanisms of our proposed method, FAQ.

D.1 Generalization Across Different Model Families

To validate that the effectiveness of FAQ extends beyond the Qwen model family, we conducted evaluations on two additional, widely-used open-source model series: Llama3 and DeepSeek.

- **Models and Data Generation**: We evaluated Llama3.1-8B-Instruct and the DeepSeek-R1 model. Consistent with our main methodology, we employed a more capable "teacher" model for synthetic data generation where

Stage	Samples	Avg. Tokens	Total Tokens	Time
Generation (Elder Sibling)				
3 candidates per query	768 (256×3)	8,240	6,328K	120 min (8×H20)
Selection (Judge)				
Scoring & Filtering	768 (256×3)	8,250	6,336K	35 min (8×H20)
FAQ Total	—	—	**12,664K**	**155 min**
PTQ (GPTQ-INT4, Qwen3-8B)	256	—	—	70 min (1×H20)

Table 12: Computational cost breakdown for FAQ on Qwen3-8B. The FAQ dataset is a one-time investment reusable across multiple target models within the same family.

Method	ARC_c	ARC_e	BoolQ	HellaS.	Lam.	OBQA	PIQA	SIQA	WinoG.	IFEval	MGSM	IncL.	Avg
BF16	53.67	82.24	85.38	59.78	78.38	36.20	80.30	42.32	70.40	73.75	59.30	54.68	64.70
INT8													
GPTQ	53.67	82.49	85.32	59.76	78.34	35.80	80.20	42.37	70.48	73.38	58.50	54.50	64.57
+FAQ	53.88	82.45	85.23	59.81	78.38	35.50	80.15	42.20	70.32	74.40	58.82	55.05	64.68
INT4													
GPTQ	50.17	81.14	83.76	58.61	77.78	33.80	79.49	41.66	69.61	70.06	55.77	52.66	62.88
+FAQ	49.49	80.72	84.68	58.97	77.06	33.20	79.16	41.97	70.40	71.35	57.70	51.56	63.02
GPTAQ	50.60	80.56	84.34	58.83	77.97	34.00	79.43	40.38	70.01	71.90	56.60	50.28	62.91
+FAQ	49.74	80.62	83.54	58.85	76.91	35.40	79.03	41.94	69.42	73.11	56.22	53.67	63.20
AVERAGE													
Quant	51.48	81.40	84.47	59.07	78.03	34.53	79.71	41.47	70.03	71.78	56.96	52.48	63.45
+FAQ	51.04	81.26	84.48	59.21	77.45	34.70	79.45	42.04	70.05	72.95	57.58	53.43	63.64

Table 13: Per-benchmark results on General tasks for Llama3.1-8B-Instruct. Higher is better.

available. For Llama3.1-8B-Instruct, data was generated by the Llama3-405B model. For DeepSeek-R1, where a significantly larger public model is unavailable, we adopted a self-generation strategy.

- **Results and Analysis**: The results for Llama3.1-8B-Instruct are presented in Table 13. Our FAQ method consistently enhances performance, particularly in the challenging INT4 setting. For example, applying FAQ to GPTAQ (GPTAQ+FAQ) improves the average accuracy to 63.20%, surpassing both the standard GPTQ baseline (62.88%) and the standalone GPTAQ (62.91%). This demonstrates that FAQ's performance benefits are robust and generalizable across different foundational model architectures.

D.2 Ablation Study on the Impact of Data Generator's Model Size

The central premise of our method, FAQ, is that leveraging high-quality synthetic data is crucial for achieving robust post-training quantization. To provide strong empirical support for this premise and to specifically investigate the role of the data generator's capability, we conducted a critical ablation study on the Qwen3-8B model. This study directly compares the quantization outcomes when using data from two distinct sources: a significantly larger "Elder Sibling" model (the Qwen3-235B), versus the model's own self-generated data. The goal is to demonstrate that the quality of the data source directly correlates with the final performance of the quantized model.

The results of this ablation, presented in Table 14, unequivocally support the core principle of FAQ. While the self-generation strategy already offers a substantial improvement over baseline quantization methods, leveraging data from the more capable "Elder Sibling" model consistently yields superior results, more effectively closing the performance gap to the unquantized BF16 baseline. This finding provides direct validation for our method's core idea: the quality and capability of the data gen-

Model	Method	AIME'24	Math-500	LiveCodeBench_v5	Avg(↑)
DeepSeek-R1	BF16	86.25	95.73	72.75	84.91
	w8a8-int8	81.67	95.38	71.26	82.77
	+FAQ	83.47	95.69	72.16	**83.77**
Qwen3-8B	BF16	83.33	95.20	58.08	78.87
	w8a8-int8	74.33	94.33	54.67	74.44
	+FAQ-8B	75.0	94.60	55.09	74.90
	+FAQ-235B	76.36	94.53	55.61	**75.5**

Table 14: Evaluation of INT8 quantization on challenging domain-specific benchmarks: AIME'24, Math-500, and LiveCodeBench (coding). This table serves two purposes: (1) It demonstrates the effectiveness of our FAQ method on the DeepSeek-R1 model. (2) It presents an ablation study on Qwen3-8B, comparing the impact of using data generated by the model itself (+FAQ-8B) versus a larger "Elder Sibling," Qwen3-235B (+FAQ-235B). The superior performance of +FAQ-235B highlights the benefit of using a more capable data generator. All values are pass@1 accuracy (%). Higher is better.

Selection Strategy	AIME-24	Math-500	LCB-v5	AIME-25	Avg
Standard GPTQ-INT4 (Baseline)	74.17	95.00	52.10	59.17	70.11
PPL Selection (Top-256 Low PPL from Raw Data)	69.17	93.81	50.34	57.29	67.65 (↓2.46)
PPL Selection + FAQ (Gen. CoT on Low PPL Seeds)	68.61	94.35	51.05	62.50	69.13 (↑1.48)
FAQ (Ours)(Random Seeds + Judge Filtering)	**73.33**	**95.10**	**51.50**	**66.67**	**71.65** (↑1.54)

Table 15: Impact of calibration data selection strategy on Qwen3-8B GPTQ-INT4. PPL-based selection degrades performance, while FAQ's regeneration recovers and surpasses the baseline. Higher is better.

erator are key determinants of post-quantization robustness. It also highlights the practical versatility of FAQ, as it remains a highly effective framework even when an "Elder Sibling" model is unavailable and a self-generation strategy must be employed.

D.3 Impact of Sample Selection Strategy

A natural question is whether simpler sample selection criteria—such as perplexity or token-length filtering—could achieve similar improvements without the regeneration overhead. To address this, we compare FAQ against several alternative selection strategies applied to the same seed calibration set: (i) *Low-PPL Selection*, selecting samples with the lowest perplexity under the target model; (ii) *High-PPL Selection*, selecting samples with the highest perplexity (representing more challenging examples); and (iii) *Random Selection*, the baseline approach.

As shown in Table 15, selecting samples with the lowest perplexity (PPL Selection) leads to a significant degradation in average accuracy (67.65 vs. 70.11 for baseline, ↓2.46), confirming that low-PPL samples are inherently "predictable" and fail to capture the complex, outlier-rich activations necessary for accurate Hessian estimation. Applying FAQ regeneration to these low-PPL seeds partially recovers performance to 69.13 (↑1.48), demonstrating that CoT generation adds valuable calibration signal even on simple seeds. The full FAQ pipeline (random seeds + Judge filtering) achieves the best overall performance (71.65), proving the necessity of both **input diversity** and **generative complexity**.

D.4 Comparison with Data-Free Calibration

To establish a performance lower bound, we evaluate a standard Data-Free quantization baseline (Per-Channel AbsMax Round-to-Nearest) on Qwen3-8B (INT8).

As shown in Table 16, even at INT8 precision, Data-Free quantization lags significantly behind our FAQ-calibrated methods. The gap is particularly pronounced on hard tasks like **AIME-24** (74.03 vs. 79.20, $\Delta = \mathbf{5.17}$), demonstrating that **activation-aware optimization** is critical for preserving mathematical reasoning capabilities. This confirms that specialized calibration data is not optional but essential—purely weight-based (data-free) methods fail to account for the model's actual activation landscape during complex reasoning, leading to substantial accuracy drops.

Method	AIME-24	Math-500	LCB-v5	AIME-25	Avg
Data-Free RTN (No Calibration Data)	74.03	94.22	53.52	65.42	71.79
GPTQ+FAQ (Ours)	79.20	95.10	58.70	62.90	73.98 (↑2.19)
GPTAQ+FAQ (Ours)	**78.30**	**94.20**	**56.30**	**69.20**	**74.50** (↑2.71)

Table 16: Impact of calibration data presence on Qwen3-8B INT8. Even at INT8 precision, Data-Free quantization lags significantly behind FAQ-calibrated methods. Higher is better.

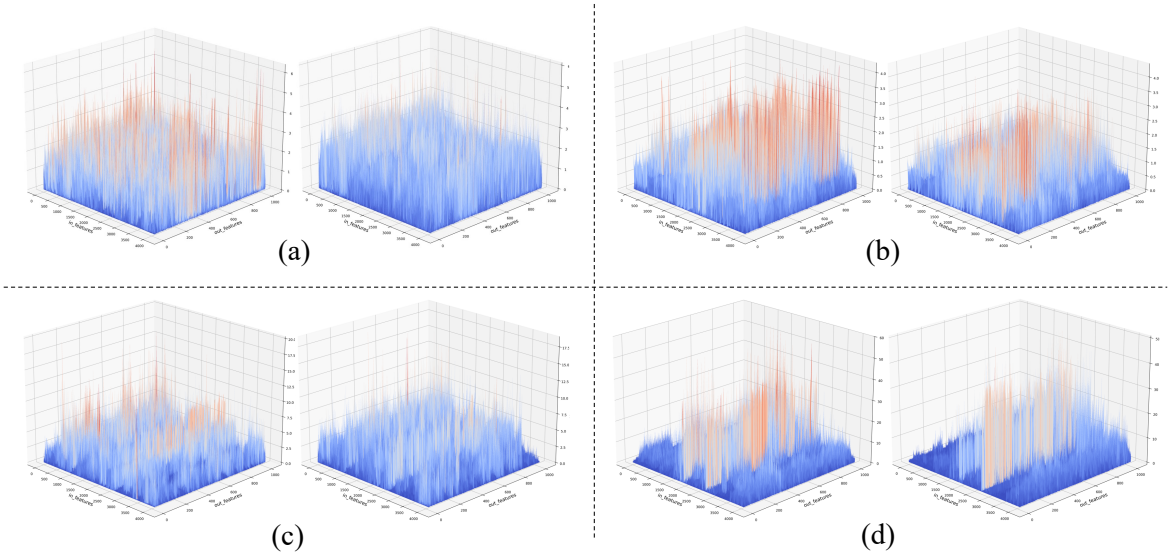
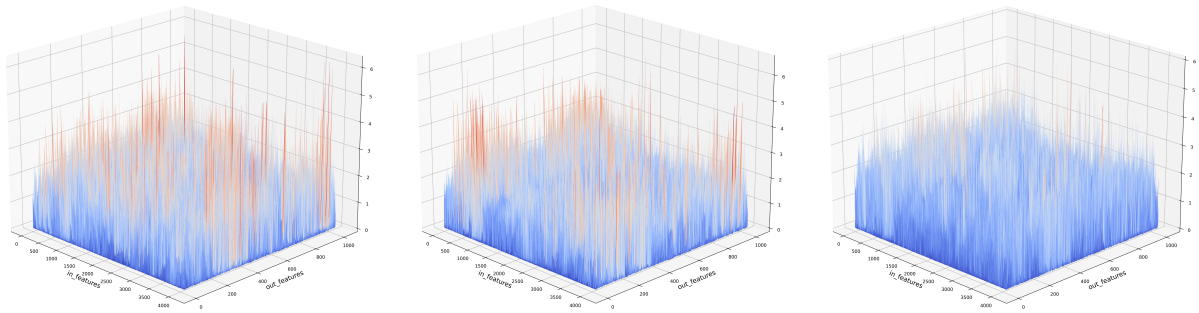


Figure 4: Activation distributions at the O_{proj} layer of Qwen3-8B. Each pair shows baseline calibration data (**left**) vs.FAQ-generated data (**right**). FAQ consistently suppresses outlier activations (red peaks), yielding a smoother, more quantization-amenable distribution.

E Visualization Analysis

To directly visualize the effect of FAQ-generated calibration data, Figure 4 presents the activation value distributions at the input of the self-attention output projection (O_{proj}) within the Qwen3-8B model across four representative layers. In each subfigure, the left landscape corresponds to the baseline calibration data, while the right corresponds to our FAQ-generated data. The contrast is consistent and striking: the FAQ-generated data produces a markedly smoother activation landscape with substantially suppressed outliers, evidenced by fewer and shorter red peaks. This directly illustrates that FAQ regeneration steers the calibration set toward a more quantization-amenable distribution.

To understand why a more capable in-family generator yields better calibration, Figure 5 visualizes the activation landscape of the *same* target model (Qwen3-8B) at the Layer-23 O_{proj}, under three different calibration sets while keeping all other PTQ configurations identical: (a) the original seed data (Figure 5a), (b) data regenerated by the target model itself (Qwen3-8B, Figure 5b), and (c) data regenerated by a larger in-family model (Qwen3-235B, Figure 5c), corresponding to our FAQ setting. The progression from (a) to (c) reveals a clear trend: the elder-sibling generator induces a noticeably smoother landscape with substantially fewer and lower outlier spikes, providing direct visual evidence that generator capacity within the same model family translates to higher-quality calibration distributions.



(a) Seed calibration data (b) Regenerated by Qwen3-8B (c) Regenerated by Qwen3-235B

Figure 5: Activation landscapes of Qwen3-8B (Layer-23 O_{proj}) under three calibration sets: (a) seed data, (b) self-regenerated (Qwen3-8B), and (c) FAQ-regenerated (Qwen3-235B). A larger in-family generator progressively suppresses outlier spikes, yielding smoother activations.

Listing 1: lm_eval:loglikelihood_rolling

```
# wikitext,c4
lm_eval --tasks wikitext,c4 \
  --model local-completions \
  --model_args max_length=16384,model=${model_name},base_url=http://${model_ip}:
      ${model_port}/v1/completions,num_concurrent=32,max_retries=3,
      tokenized_requests=False,timeout=72000 \
  --num_fewshot 0 \
  --gen_kwargs temperature=0.6,max_gen_toks=32768,top_k=20,top_p=0.95
```

Listing 2: lm_eval:generate_until

```
# ifeval
lm_eval --tasks ifeval \
  --model local-chat-completions \
  --model_args max_length=32768,model=${model_name},base_url=http://${model_ip}:
      ${model_port}/v1/chat/completions,num_concurrent=32,max_retries=3,
      tokenized_requests=False,timeout=72000 \
  --num_fewshot 0 \
  --apply_chat_template qwen3 \
  --gen_kwargs temperature=0.6,max_gen_toks=32768,top_k=20,top_p=0.95
```

Listing 3: evalscope

```

# AIME2024,Aime2025,MATH500,LiveCodeBench
evalscope eval --datasets aime24 aime25 math_500 live_code_bench \
  --model ${model_name} \
  --api-url http://${host}:${port}/v1/chat/completions \
  --api-key EMPTY \
  --eval-type service \
  --eval-batch-size 4 \
  --generation-config '{"do_sample":true,"temperature":0.6,"top_p":0.95,"top_k
      ":20,"max_tokens":32768,"n":8,"chat_template_kwargs":{"enable_thinking":
      true}}' \
  --chat-template qwen3 \
  --stream \
```

Listing 4: sglang

```
python -m sglang.launch_server \
    --model-path models/Qwen3-30B-A3B --trust-remote-code \
    --served-model-name Qwen3-30B-A3B --port ${port} \
    --tensor-parallel-size 4 --mem-fraction-static 0.9 --attention-backend fa3 \
    --reasoning-parser qwen3 --enable-torch-compile \
    --torch-compile-max-bs 32 --cuda-graph-max-bs 32
```

Listing 5: vllm

```
python3 -m vllm.entrypoints.openai.api_server \
    --model /models/Qwen3-8B --trust-remote-code \
    --served-model-name Qwen3-8B --port ${port} \
    --tensor-parallel-size 4 --gpu-memory-utilization 0.85 \
    --max-model-len 34816 --reasoning-parser qwen3 \
    --max-num-seqs 32
```