Target-Aware Calibration Data Selection for Preserving Uncertainty in Quantized Language Models

Zhen Yang[1,†], **Sizai Hou**[2,†], **Kaiwen Zheng**[3], **Yaofang Liu**[4], **Liang He**[5]
Yixuan Chen[6,*], **Kangning Cui**[7,*]

[1]Yale University [2]The Hong Kong University of Science and Technology
[3]The Hong Kong University of Science and Technology (Guangzhou)
[4]City University of Hong Kong [5]Shanghai Institute of Optics and Fine Mechanics
[6]University of Oxford [7]City University of Hong Kong (Dongguan)
[†] Equal contribution [*] Corresponding authors

Abstract

Quantization is widely used to deploy large language models, but its effect on uncertainty behavior, such as confidence, margins, and abstention, is rarely treated as a primary objective. We frame calibration-data selection for quantization as a *target-dependent uncertainty-preservation problem*. Different deployments emphasize different regions of the input distribution, yet prior work mainly optimizes accuracy-oriented compression metrics or adjusts scores after quantization. We formalize this goal with distributional and boundary preservation risks, and provide a simple mixture-mismatch argument explaining why no single calibration recipe should be expected to fit all targets. We introduce Doubt-Preserving Quantization (DPQ), a lightweight pre-quantization recipe family that uses full-precision predictions to construct target-aligned calibration mixtures of high-doubt examples and generic anchors. Across 8 language models, 9 NLP benchmarks, and 22 comparison methods, the leading fixed recipe changes with the preservation target: DPQ-r75 leads on SQuAD2 answerability-boundary preservation, while milder or single-signal variants, including DPQ-r50, confidence-only, and entropy-only, better preserve broad multiple-choice QA behavior. These results show that calibration data should be selected for the specific full-precision score behavior a deployment needs to preserve, rather than treated as a fixed quantization detail. Code is available at https://github.com/xi-xiaoran/DPQ.

1 Introduction

Quantization has become a standard tool for deploying large language models (LLMs) under memory and compute budgets. Weight-only and low-bit post-training methods such as GPTQ (Frantar et al., 2023), AWQ (Lin et al., 2024), SmoothQuant (Xiao et al., 2023), OmniQuant (Shao et al., 2024), LLM.INT8() (Dettmers et al., 2022), SpQR (Dettmers et al., 2024), and QLoRA-style 4-bit representations (Dettmers et al., 2023) make large models accessible on smaller hardware. Most compression evaluations, however, still focus on accuracy, perplexity, or benchmark scores (He et al., 2026; Wu et al., 2026). This is incomplete for applications that rely on model scores for abstention, reranking, ensembling, verification, or answerability decisions. In these settings, preserving confidence, margins, and answerability decisions can be as important as preserving the top-1 answer (Shao et al., 2026). Recent work shows that quantization can shift confidence (Proskurina et al., 2024) and that calibration data affects compression quality (Williams and Aletras, 2024), but calibration-data selection rarely targets preservation of FP uncertainty behavior.

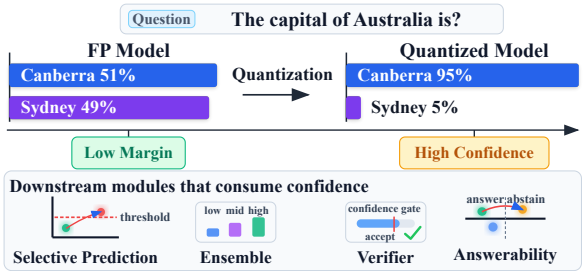

Figure 1: Motivating example of uncertainty drift after quantization. The top-1 answer is unchanged, but the confidence margin shifts substantially, affecting downstream uncertainty-aware decisions.

This distinction matters even when accuracy is unchanged. As illustrated in Figure 1, an FP model may assign option probabilities $(0.51, 0.49)$ while its quantized counterpart assigns $(0.95, 0.05)$, preserving the same top-1 answer but changing the confidence margin. Such shifts can alter downstream uncertainty-aware decisions, especially in answerability detection, where a small score change may determine whether the system answers an unanswerable question. Existing ap-

proaches do not directly target this preservation problem. Accuracy-oriented calibration-data studies (Williams and Aletras, 2024) optimize reconstruction quality or label accuracy, while post-hoc calibration (Guo et al., 2017; Zhong et al., 2025) adjusts scores after quantization. Temperature-style maps preserve the argmax and cannot repair decision flips, whereas more flexible maps can improve label calibration at the cost of moving the quantized model away from FP behavior.

We therefore frame calibration-data selection for quantization as a target-dependent *uncertainty-preservation problem*. Broad answerable tasks require distributional preservation of option probabilities, confidence, and margins, whereas answerability or abstention-heavy tasks require boundary preservation of low-margin answer/no-answer decisions. We formalize these targets as two preservation risks and give a mixture-mismatch argument that explains why no fixed mixture of boundary examples and generic anchors should be expected to fit all targets unless the targets coincide. Building on this result, we introduce DPQ (Doubt-Preserving Quantization), a lightweight pre-quantization framework that selects high-doubt and answerability-boundary examples using FP predictions, mixes them with generic anchors, and runs the unchanged quantizer. Across 8 language models, 9 NLP benchmarks, and 22 comparison methods, we observe this target dependence: DPQ-s128-r75 performs best for SQuAD2 answerability-boundary preservation, while broad MCQA favors milder or single-signal recipes such as DPQ-r50, confidence-only, and entropy-only. Post-hoc and AWQ analyses further show that preservation signals are partly transferable, but score-based metrics remain quantizer-specific.

Contributions.

- **Problem.** We formulate calibration-data selection for quantization as preserving full-precision uncertainty behavior, with distributional and boundary risks for different deployment targets.
- **Method.** We provide a mixture-mismatch argument and introduce DPQ as a lightweight target-aware recipe family: it uses FP predictions to select high-doubt or boundary-near examples, mixes them with generic anchors, and leaves the quantizer unchanged.
- **Evaluation.** Across 8 language models, 9 NLP benchmarks, and 22 comparison methods, we show that the best recipe depends on the preservation target: high-boundary mixtures better preserve answerability boundaries, while milder recipes better preserve broad MCQA behavior.

2 Related Work

LLM quantization. Post-training LLM quantization includes approximate second-order weight quantization (Frantar et al., 2023), activation-aware and activation-smoothing methods (Lin et al., 2024; Xiao et al., 2023), 8-bit and 4-bit representations (Dettmers et al., 2022, 2023), and recent low-bit PTQ systems (Yao et al., 2022; Shao et al., 2024; Dettmers et al., 2024; Kim et al., 2024; Egiazarian et al., 2024; Tseng et al., 2024; Ashkboos et al., 2024). These methods primarily optimize reconstruction or downstream accuracy. Calibration data supports quantization, but is less often selected with the explicit goal of preserving confidence, margins, or abstention behavior.

Calibration data and selection. Williams and Aletras (2024) show that calibration data affects pruning and quantization, while self-calibration uses the model itself to generate calibration data for quantization and pruning (Williams et al., 2025). We study a different objective: selecting calibration data to preserve full-precision uncertainty behavior. Related data-selection methods, including uncertainty sampling, core-set selection, and gradient- or representation-diversity selection, have been studied for data efficiency (Settles, 2009; Sener and Savarese, 2018; Ash et al., 2020). These objectives are useful controls, but they do not directly target quantized-vs-FP option-probability preservation; we therefore include them as baselines.

Confidence calibration. Calibration has a long history in probabilistic prediction and evaluation (Brier, 1950; Niculescu-Mizil and Caruana, 2005; Guo et al., 2017; Nixon et al., 2019; Platt, 1999; Zadrozny and Elkan, 2002; Naeini et al., 2015; Kull et al., 2019; Kumar et al., 2019; Minderer et al., 2021). Selective prediction and uncertainty estimation have also been widely studied (Chow, 1970; El-Yaniv and Wiener, 2010; Geifman and El-Yaniv, 2017; Ovadia et al., 2019). For language models, prior work studies confidence, self-knowledge, calibration, and "know-when-you-do-not-know" behavior in QA, prompting, and tuning settings (Jiang et al., 2020; Zhao et al., 2021; Desai and Durrett, 2020; Xiao et al., 2022; Kadavath et al., 2022;

Kapoor et al., 2024a,b). These works mainly analyze FP models, prompting, or direct tuning, rather than calibration-data choice during quantization.

Post-hoc calibration. Proskurina et al. (2024) document confidence shifts under low-bit compression, and Zhong et al. (2025) propose soft-prompt post-hoc calibration. These works intervene after quantization is fixed. Temperature-style maps preserve the argmax and cannot repair decision or answerability flips; flexible score-space calibrators can change decisions but may move the model away from FP behavior (Guo et al., 2017; Zadrozny and Elkan, 2002; Naeini et al., 2015; Kull et al., 2019). Our work instead studies calibration-data selection before quantization.

3 Problem Statement

We formulate calibration-data selection for quantization as uncertainty preservation with a task-dependent target. §3.1 defines distributional and boundary risks. §3.2 explains why low-margin examples are fragile under quantization perturbations. §3.3 gives a mixture-mismatch argument that motivates the DPQ design space in §4.

3.1 Preservation Risks

Option scoring. For an input x with candidate options $\mathcal{Y}(x) = \{y_1, \dots, y_K\}$, a full-precision model assigns a score $s_0(x, y_i)$ to each option and induces

$$p_0(y_i \mid x) = \frac{\exp s_0(x, y_i)}{\sum_{j=1}^{K} \exp s_0(x, y_j)}, \qquad (1)$$

where s_0 is computed from conditional language-model likelihood. A quantized model produced with calibration set $\mathcal{D}$ induces scores $s^Q_{\mathcal{D}}(x, y_i)$ and probabilities $p^Q_{\mathcal{D}}(y_i \mid x)$ analogously. Standard accuracy checks whether the top option matches the label; we instead ask whether $p^Q_{\mathcal{D}}$ preserves the *uncertainty behavior* of p_0. In SQuAD2, answerability is scored as a two-option choice, and $p(\text{answerable})$ denotes the softmax probability of the answerable option.

Risks. Expectations are over the target evaluation distribution. We define two preservation risks. The first is

$$\mathcal{R}_{\text{dist}}(\mathcal{D}) = \mathbb{E}_x\big[\text{JSD}(p^Q_{\mathcal{D}}(\cdot \mid x), p_0(\cdot \mid x))\big], \quad (2)$$

which measures quantization-induced change in the option distribution. The second is

$$\mathcal{R}_{\text{bdry}}(\mathcal{D}) = \mathbb{E}_x\big[\mathbf{1}\{a^Q_{\mathcal{D}}(x) \neq a_0(x)\}\, w(x)\big], \quad (3)$$

where $a_0(x) = \arg\max_y p_0(y \mid x)$ and $a^Q_{\mathcal{D}}(x) = \arg\max_y p^Q_{\mathcal{D}}(y \mid x)$ are the FP and quantized decisions, and $w(x) \geq 0$ upweights answerability or low-margin cases. $\mathcal{R}_{\text{dist}}$ matters when the quantized model should act as a drop-in replacement for an FP model, as in broad MCQA, reranking, or ensembling. $\mathcal{R}_{\text{bdry}}$ matters when answerability or abstention decisions are central, as in SQuAD2 answerability, safety filters, or selective prediction.

Metrics. Empirically, we instantiate these risks with ECE (Guo et al., 2017), adaptive ECE (Nixon et al., 2019), NLL, and Brier score (Brier, 1950), together with FP-behavior metrics: FP agreement, $\text{JSD}(p^Q_{\mathcal{D}}, p_0)$, confidence shift, top-two margin drift, and margin correlation. For SQuAD2, we also report boundary-accuracy deviation and answerability-rate deviation, which measure how far the quantized answer/abstain decisions and answer rate move from FP. These metrics evaluate fidelity to the full-precision reference rather than improved ground-truth calibration: FP agreement and JSD-to-FP measure whether a quantized model can replace an already validated FP model in downstream score-consuming pipelines.

3.2 Boundary Fragility

Uncertainty behavior is most fragile near decision boundaries. Let $s_0(x) \in \mathbb{R}^K$ and $s^Q_{\mathcal{D}}(x) \in \mathbb{R}^K$ be the FP and quantized option-score vectors, and define $\Delta_{\mathcal{D}}(x) = s^Q_{\mathcal{D}}(x) - s_0(x)$. Let i^* and j^* be the FP top and runner-up options, with margin $\gamma(x) = s_0(x, i^*) - s_0(x, j^*) > 0$.

Proposition 1 (Top-two boundary fragility)**.** *If $\Delta_{\mathcal{D}}(x, j^*) - \Delta_{\mathcal{D}}(x, i^*) > \gamma(x)$, then the relative ordering of i^* and j^* is reversed by quantization. Conversely, if $\|\Delta_{\mathcal{D}}(x)\|_\infty < \gamma(x)/2$, the* FP *top-1 decision is preserved.*

Thus, small-margin examples can flip under small perturbations, while large-margin examples are stable. This does not imply that GPTQ directly optimizes option scores; it shows where reconstruction error becomes *behavioral* error. Generic-text calibration may reduce average reconstruction error while missing activation directions that control answerability and confidence. Boundary-aware calibration exposes these fragile directions to the quantizer.

3.3 Mixture Mismatch

The two risks emphasize different parts of the input space. We make this explicit with a mixture view.

Mixture view. Let q_{bdry} denote the distribution over boundary or high-doubt examples, and let q_{anchor} denote the distribution over generic anchors, such as WikiText or random QA. A calibration recipe is $q_r = r\, q_{\text{bdry}} + (1-r)\, q_{\text{anchor}}$, where $r \in [0,1]$ is the boundary ratio used by DPQ; for example, DPQ-r75 corresponds to r=0.75. The target preservation distribution is $T_\theta = \theta q_{\text{bdry}} + (1-\theta) q_{\text{anchor}}$. Answerability-boundary preservation has large θ, while broad answerable MCQA preservation has small θ.

Fixed recipes. Generic-text calibration, such as WikiText or C4, corresponds roughly to $r \approx 0$. Boundary-only and answerability-only variants correspond to $r \approx 1$. Hard-example mining selects a different region: an example can be confidently wrong, with high NLL and large margin, without being uncertain; hence high-NLL $\neq$ high-doubt. Post-hoc calibration intervenes at a different stage: temperature maps preserve argmax and cannot repair flips, while flexible maps can change decisions but may move the model away from FP behavior. These fixed choices cannot match every target; formal statements appear in Appendix C.

Proposition 2 (Mixture-ratio mismatch)**.** *Let $\ell_{\mathcal{D}}(x) \in [0,1]$ be a bounded preservation loss for calibration set $\mathcal{D}$, and define $R_P(\mathcal{D}) = \mathbb{E}_{x\sim P}[\ell_{\mathcal{D}}(x)]$. For any two distributions T and q,*

$$|R_T(\mathcal{D}) - R_q(\mathcal{D})| \le \mathrm{TV}(T, q).$$

If q_{bdry} and q_{anchor} have disjoint support, then

$$\mathrm{TV}(T_\theta, q_r) = |\theta - r|, \tag{4}$$

so the mismatch upper bound is minimized at $r^* = \theta$. Consequently, if two targets have different $\theta_1 \neq \theta_2$, no single r minimizes mismatch to both.

The proof and behavioral-risk surrogate appear in Appendix C.3. This result should be read as a design principle rather than an exact predictor of the best ratio for a given quantizer. It motivates evaluating target-dependent mixtures and interpreting the empirical winner as a target–recipe match, rather than expecting one fixed recipe to dominate all targets. DPQ provides this design space, and the experiments test the target–recipe match.

4 Method

DPQ is a pre-quantization calibration-data selection strategy for GPTQ-style post-training quantization. Algorithm 1 gives the procedure. DPQ does not modify the quantizer kernel, bit-width, reconstruction objective, or inference path; it only changes the calibration strings used to estimate activation statistics. We describe the method through the calibration-selection view in §4.1 and the doubt-based criterion in §4.2.

Algorithm 1 DPQ Calibration Data Selection

Input: FP model M; candidate pool C; calibration size s; mixture ratio r
Output: Calibration set $\mathcal{D}$ with $|\mathcal{D}| = s$

1: Score each $x \in C$ with M to obtain $p_0(\cdot \mid x)$
2: Compute $b(x) = 1 - \big(p_{0,(1)}(x) - p_{0,(2)}(x)\big)$ for each $x \in C$
3: Select C_{hi} as the top $\lfloor sr \rfloor$ candidates by $b(x)$, with answerability balancing for SQuAD2 boundary data
4: Draw C_{anc} with $s - \lfloor sr \rfloor$ anchors from WikiText or RandomQA
5: Build D_{hi} from C_{hi} using the original prompt and FP top option or options
6: Set $\mathcal{D} \leftarrow D_{\text{hi}} \cup C_{\text{anc}}$
7: Run unchanged GPTQ with calibration set $\mathcal{D}$

4.1 Calibration Selection

GPTQ estimates layer-wise activation statistics from a calibration set $\mathcal{D}$ to solve a local reconstruction problem, so $\mathcal{D}$ determines which activation regions are represented accurately. We view calibration selection as choosing $\mathcal{D}$ from a distribution q that approximates the target preservation distribution T_θ from §3.3. Proposition 2 motivates allocating calibration mass according to the target mixture over boundary and anchor components.

Our candidate pool C combines training-split ARC-Challenge examples, which provide answerable QA structure, with SQuAD2 answerability examples, which provide balanced answerable and unanswerable boundary cases. In our implementation, C contains 512 candidates from each source before model-specific scoring. The FP model scores C once under the evaluation option-scoring protocol, producing $\{p_0(\cdot \mid x)\}_{x\in C}$; this is an offline selection cost and adds no inference-time overhead. Evaluation examples are never used for calibration, and the same pool is reused across all DPQ variants. This setup reflects the deployment view of calibration selection: the pool provides target-relevant candidates, while DPQ determines which examples within that pool best preserve the desired FP behavior.

4.2 Doubt-Based Selection

For a candidate x, let $p_{0,(1)}(x)$ and $p_{0,(2)}(x)$ denote the largest and second-largest FP option probabilities. We define the doubt score

$$b(x) = 1 - \big(p_{0,(1)}(x) - p_{0,(2)}(x)\big), \qquad (5)$$

which assigns high values to low-margin examples, the fragile region identified by Proposition 1. To isolate the effect of each signal, we also evaluate component variants that replace $b(x)$ with $1-p_{0,(1)}(x)$ for confidence-only, $H(p_0(\cdot \mid x))$ for entropy-only, answerability log-odds for answerability-only, or gold-label NLL for hard-example mining.

Given budget s and mixture ratio $r \in [0,1]$, DPQ selects the top $\lfloor sr \rfloor$ candidates by $b(x)$ as boundary strings and draws the remaining $s - \lfloor sr \rfloor$ anchor strings from WikiText-style text and RandomQA-style calibration. Each boundary string concatenates the original prompt with the FP top option or options, while anchors are kept unchanged. This construction is fixed across DPQ variants, so the ablations compare selection signal, ratio, and budget under the same calibration-string design. The ratio r is the mixture weight from §3.3: higher r targets boundary-heavy settings such as SQuAD2 answerability, and lower r targets broader answerable behavior. We evaluate ratio, size, boundary, and component variants as controlled DPQ recipes in Section 5. Because DPQ only changes calibration strings, the resulting model has the same form as a standard GPTQ baseline and incurs no additional inference cost. Thus, DPQ is best viewed as a target-aware recipe family rather than a single universal calibration set.

5 Experiment Results

5.1 Setup

Evaluation scope. We evaluate 8 LMs across three model families and multiple scales: Qwen2.5-0.5B/1.5B/3B/7B (Qwen Team, 2024); Llama-3.2-1B, Llama-3.2-3B (Meta AI, 2024), and Llama-3.1-8B (Grattafiori et al., 2024); and Mistral-7B-v0.3 (Jiang et al., 2023). The benchmark suite contains 9 NLP datasets organized by preservation target. The old-core suite includes ARC-Challenge (Clark et al., 2018), SQuAD2 answerability (Rajpurkar et al., 2018), and TruthfulQA (Lin et al., 2022). Within this suite, SQuAD2 is the primary boundary-preservation benchmark because it directly tests answerable versus unanswerable decisions, while ARC-Challenge and TruthfulQA provide additional checks of score behavior. The extra-MCQA suite includes ARC-Easy (Clark et al., 2018), BoolQ (Clark et al., 2019), PIQA (Bisk et al., 2020), HellaSwag (Zellers et al., 2019), OpenBookQA (Mihaylov et al., 2018), and CommonsenseQA (Talmor et al., 2019); since these examples are answerable, they primarily test broad distributional preservation.

Method	Worse than FP (%, ↓)			
	Acc.	ECE	NLL	Brier
BNB-NF4	84.7	66.7	84.7	81.9
GPTQ-RandomQA	83.3	61.1	76.4	86.1
GPTQ-WikiText	76.4	58.3	63.9	83.3
GPTQ-C4	83.3	65.3	68.1	77.8
DPQ-s128-r75	80.6	62.5	70.8	83.3
DPQ-s128-r50	79.2	58.3	65.3	80.6

Table 1: Low-bit quantization often worsens accuracy and calibration metrics. Entries show the percentage of model–dataset settings worse than FP; lower is better. Full results are in Table 11.

Methods and ranking. For each model–dataset pair, the main comparison includes one full-precision reference and 22 comparison methods: BNB-NF4; generic and task-formatted GPTQ calibration such as WikiText, C4, RandomQA, and TaskRandom; DPQ ratio, size, boundary, and component variants; data-selection baselines; negative controls; and post-hoc controls. All main quantized baselines use 4-bit quantization; BNB-NF4 is included as an NF4 quantizer-family baseline, while the GPTQ variants differ only in calibration data. Full method inventory, prompt formats, and quantization hyperparameters are provided in Appendix D. We rank methods using target-aligned metrics: for SQuAD2 boundary preservation, boundary-accuracy deviation, answerability-rate deviation, FP agreement, and JSD; for broad MCQA preservation, FP agreement, JSD, margin drift, confidence shift, accuracy deviation, and margin correlation. Within each target, metrics are equally weighted as a neutral summary because no deployment-specific utility function is assumed. Ranks are averaged over the 8 models and used only as compact summaries; our interpretations rely on component metrics and target-specific trends rather than on rank alone. Raw metrics and per-dataset breakdowns are in Appendix A.

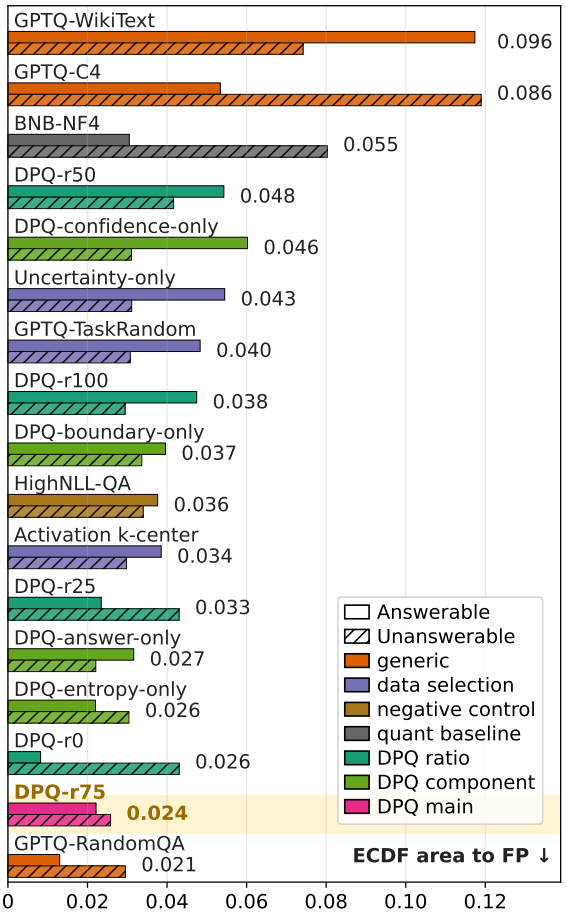

Figure 2: SQuAD2 answerability ECDF area to FP per pre-quantization method, averaged over Qwen2.5-7B, Llama-3.1-8B, and Mistral-7B-v0.3. Numbers denote the mean of the answerable and unanswerable areas. Methods are sorted top-to-bottom by mean area (worst first); DPQ-r75, the best aggregate boundary recipe in Table 2, is highlighted.

5.2 Quantization-Induced Drift

Table 1 establishes the empirical problem. Across representative quantization and calibration choices, degradation is frequent not only in accuracy but also in calibration-sensitive metrics such as ECE, NLL, and Brier score. Thus, low-bit quantization often changes score behavior, not just final answers. The full degradation table, including confidence shift and margin drift, is provided in Table 11. In particular, Table 12 in Appendix B.1 shows that quantization compresses the confidence separation between correct and incorrect cases, although the direction of absolute confidence shifts is task-dependent.

This motivates the target-specific analysis that follows: *given widespread quantization-induced drift, which calibration data choices best preserve the uncertainty behavior required by each deployment target?* We answer this separately for answerability-boundary preservation and broad MCQA preservation.

5.3 Answerability Boundary

Table 2 reports the core SQuAD2 answerability-boundary result. DPQ-s128-r75 is the strongest listed pre-quantization recipe, with the best aggregate rank, highest agreement with FP, and smallest boundary-accuracy deviation, answerability-rate deviation, and JSD. Compared with GPTQ-WikiText, it reduces boundary-accuracy deviation from 0.4553 to 0.2125, answerability-rate deviation from 0.2271 to 0.1000, and JSD from 0.0387 to 0.0158. The conclusion does not rely on rank alone: the same recipe also improves the absolute boundary and distributional metrics. Figure 2 gives a complementary deployment-scale view, showing that DPQ-r75 is among the recipes closest to the FP answerability distribution. Together, these results show that boundary-heavy targets benefit from high-doubt calibration examples, while the mixed r75 recipe also indicates that generic anchors are needed. The ordering remains stable after removing the Llama-3.2-1B stress case; see Table 6.

Distributional mechanism. Figure 3 gives a distributional view of the SQuAD2 result on Llama-3.1-8B. For gold-answerable examples, $p(\text{answerable})$ should remain near one; for gold-unanswerable examples, it should remain near zero. Generic GPTQ-WikiText calibration shifts the answerability distribution away from FP, especially on answerable examples, while DPQ-r75 moves it closer to the FP reference. This illustrates why SQuAD2 answerability is not simply a harder MCQA setting: an answerability flip changes whether the system should answer at all. The pattern is consistent with using a boundary-heavy mixture for this target, while the ablations below show that anchors are still needed for stability. Additional ECDF diagnostics for Qwen2.5-7B and Mistral-7B-v0.3 show the same qualitative pattern in Appendix A.1.

Ablation summary. The ablations in Appendix A.1 test ratio, size, boundary mixing, single-signal components, and negative controls. As summarized in Table 3, the gains mainly come from calibration-set composition rather than budget size: r75 performs best, r100 over-concentrates on boundary cases, and boundary-only or single-signal

Method	Rank ↓	Boundary preservation metrics			
		Agr. ↑	AccΔ ↓	RateΔ ↓	JSD ↓
DPQ-s128-r75	**6.20**	**0.8495**	**0.2125**	**0.1000**	**0.0158**
DPQ-boundary-random	8.15	0.8413	0.2310	0.1135	0.0174
DPQ-answerability-only	8.18	0.8286	0.2592	0.1281	0.0205
DPQ-entropy-only	8.68	0.8270	0.2940	0.1470	0.0227
DPQ-boundary-only	9.06	0.8335	0.2520	0.1242	0.0197
Uncertainty-only	9.36	0.8261	0.2937	0.1461	0.0181
Activation k-center	9.39	0.8360	0.2530	0.1255	0.0187
DPQ-s128-r0	10.31	0.8097	0.3205	0.1598	0.0220
GPTQ-RandomQA	11.09	0.8266	0.2733	0.1351	0.0231
DPQ-s128-r100	12.44	0.8080	0.3340	0.1655	0.0215
DPQ-s128-r50	13.12	0.8140	0.3210	0.1602	0.0258
GPTQ-WikiText	17.29	0.7446	0.4553	0.2271	0.0387

Table 2: SQuAD2 answerability-boundary preservation. Rank aggregates agreement with FP (Agr.), boundary-accuracy deviation (AccΔ), answerability-rate deviation (RateΔ), and JSD over 8 models; lower is better except Agr. Full ablations are in Table 7.

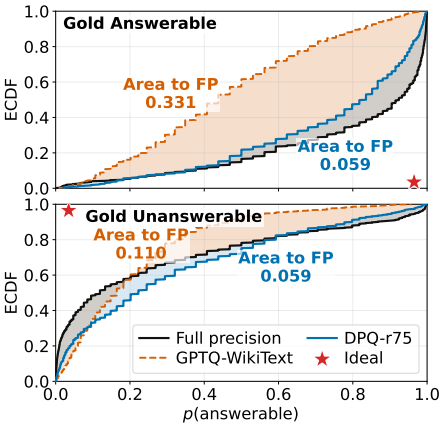

Figure 3: SQuAD2 answerability ECDFs for Llama-3.1-8B. Shaded regions show area to FP; smaller is better. Red stars mark ideal behavior.

Check	Main finding
Boundary ratio r	DPQ-r75 performs best. A high boundary ratio helps, but using only boundary examples over-concentrates the calibration set.
Calibration size s	The composition of the calibration set matters more than simply increasing the calibration budget.
Boundary mixing	Mixed calibration outperforms boundary-only variants, suggesting that generic anchors stabilize boundary-focused selection.
Single-signal variants	Confidence-only, entropy-only, and answerability-only variants are useful, but none matches the mixed DPQ-r75 recipe.
Negative controls	HighNLL-QA and LowDoubt-QA underperform DPQ-r75, showing that high doubt is different from ordinary hard- or easy-example selection.

Table 3: Ablation summary for the SQuAD2 boundary target. Full results are in Appendix A.1.

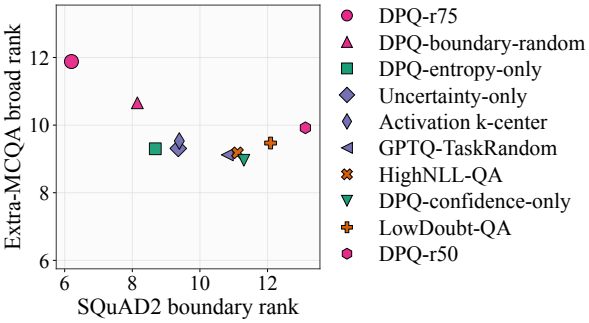

Figure 4: Target-dependent preservation trade-off: SQuAD2 boundary rank versus extra-MCQA broad-preservation rank. Lower is better on both axes; no calibration recipe dominates both targets.

variants do not match the mixed recipe. Strong data-selection baselines improve over generic text calibration but still fall below DPQ-r75 on the boundary-specific aggregate. These ablations test the main selection axes of the calibration set; the boundary-string construction is held fixed as part of the DPQ recipe.

5.4 Broad MCQA Trade-off

The six extra-MCQA benchmarks contain only answerable examples, so the relevant target is broad FP-behavior preservation ($\mathcal{R}_{\text{dist}}$), not answerability-boundary preservation. Table 4 shows that the leading recipes therefore shift away from the high-boundary setting: confidence-only, task-random, HighNLL-QA, entropy-only, uncertainty-only, and DPQ-r50 are more competitive than DPQ-r75. This pattern matches the mixture view: when the target places less mass on answerability boundaries, a smaller boundary ratio or a single uncertainty signal can better preserve broad option-distribution behavior. Figure 4 visualizes the same target dependence: DPQ-r75 is stronger on the SQuAD2 boundary axis, whereas milder or single-signal recipes are stronger on broad MCQA. The metric columns also show why Table 4 should be read as a trade-off rather than a single scalar leaderboard: BNB-NF4 has the strongest agreement and JSD, whereas other recipes rank higher under the full multi-metric objective. Expanded extra-MCQA results in Appendix A.2 support the target-dependent recipe shift, while the calibration-gap and correct-vs-wrong confidence diagnostics in Appendix B.1 further show that quantization can alter score behavior in ways not captured by top-1 accuracy alone.

Method	Rank ↓	Top-5 ↑	Agr. ↑	JSD ↓
DPQ-confidence-only	8.96	7	0.8357	0.0284
GPTQ-TaskRandom	9.12	11	0.8361	0.0276
HighNLL-QA	9.18	6	0.8327	0.0294
DPQ-entropy-only	9.30	10	0.8324	0.0296
Uncertainty-only	9.31	4	0.8318	0.0297
LowDoubt-QA	9.47	9	0.8328	0.0286
Activation k-center	9.53	12	0.8349	0.0299
DPQ-s128-r50	9.92	4	0.8322	0.0293
BNB-NF4	10.53	12	0.8637	0.0241
DPQ-s128-r75	11.88	3	0.8180	0.0324

Table 4: Broad MCQA preservation on six answerable datasets. Rank aggregates agreement with FP (Agr.), JSD, margin drift, confidence shift, accuracy deviation, and margin correlation; Top-5 counts model–dataset pairs where the method ranks in the top five. Milder, single-signal, or broad-selection recipes lead, unlike on the SQuAD2 boundary target.

Method	Acc. ↑	Agr. ↑	JSD ↓	MarginΔ ↓
Base	0.6833	0.8245	0.0320	0.1533
Adaptive temp.	0.6833	0.8245	0.0445	0.2385
Option bias	0.6985	0.8147	0.0328	0.1591
Vector	0.7029	0.7773	0.0519	0.2344
Matrix	0.6992	0.7723	0.0550	0.2380
Dirichlet	0.6989	0.7740	0.0548	0.2400
Isotonic	0.6991	0.7710	0.0561	0.2329

Table 5: Post-hoc calibration on extra-MCQA. Flexible calibrators improve accuracy but reduce agreement with FP and increase distributional drift.

Practical takeaway. In deployment, the recipe should be chosen by the preservation metric that matches the downstream use. If validation examples reflect answerability, abstention, or other low-margin decisions, a boundary-heavy mixed recipe such as DPQ-r75 is a strong choice. If the target mainly uses broad option scores on answerable MCQA tasks, milder or single-signal recipes are more appropriate. Thus, DPQ is best used as a target-aware recipe family, not as a universal calibration set selected by task accuracy alone.

Negative controls. The negative controls further clarify the target shift. HighNLL-QA and LowDoubt-QA are competitive on broad MCQA but not on SQuAD2, matching the distinction between difficulty and doubt formalized in Appendix C.4. Thus, ordinary hard or easy examples can overlap with broad answerable uncertainty, but they are not answerability-boundary examples.

Post-hoc calibration targets a different objective. Post-hoc calibration targets label-oriented calibration or task accuracy, whereas our objective is preserving FP score behavior; Table 5 illustrates this distinction. Adaptive temperature leaves accuracy and agreement with FP unchanged, as expected from Proposition 5, but increases JSD and margin drift. Flexible score-space calibrators improve MCQA accuracy, yet reduce agreement with FP and increase distributional drift. Thus, post-hoc calibration and pre-quantization selection are complementary intervention points: the former reshapes scores after quantization, while the latter controls which FP score behavior the quantized model tends to preserve. The optimally fitted temperature check in Appendix B.2 supports the same complementary pattern.

AWQ transfer. The AWQ extension serves as a quantizer-family scope check. On 7B/8B-scale models, DPQ-r75 leads on AWQ mean rank, top-1 frequency, and agreement with FP; its mean rank is 1.97, with 42.9% top-1 and 73.0% top-2 frequency. However, the per-dataset breakdown in Table 10 shows that score-based metrics can favor generic calibration on some datasets. Thus, the target-aware lens partially transfers, but the exact recipe should be tuned to the quantizer and metric; the per-dataset breakdown is in Appendix A.4 and aggregate summaries are in Appendix B.3.

6 Conclusion

Quantization should not be evaluated only by whether it preserves a model's top-1 answer. In applications that rely on confidence, margins, selective prediction, abstention, or reranking, a quantized model may keep the same prediction while changing the uncertainty behavior used downstream. We study calibration-data selection as a target-dependent uncertainty-preservation problem and show that different targets favor different recipes: high-boundary mixtures such as DPQ-r75 better preserve SQuAD2-style answerability boundaries, while milder mixtures or single-signal variants better preserve broad MCQA behavior. The central insight is therefore not that one calibration set is universally best, but that calibration data should be selected for the specific FP score behavior a deployment needs to preserve. More broadly, pre-quantization data selection, post-quantization calibration, and quantizer choice affect preservation differently, so recipes should be tuned to both the target behavior and the quantization method.

Limitations

Option-scoring uncertainty. We operationalize uncertainty through option scoring and answerability. This controlled setting allows confidence, margins, and full-precision agreement to be measured directly, and it covers common uses such as multiple-choice QA, selective prediction, abstention, and reranking. Extending the same preservation perspective to verbalized confidence and long-form generation is an important direction for future work, since uncertainty in those settings may appear in the generated text rather than in option probabilities.

Target-oriented candidate pool. Our candidate pool is intentionally target-oriented: it includes answerable QA examples and answerability-style training cases because the main boundary target is answerable/unanswerable preservation. This design matches the goal of studying deployment-aware calibration selection, where the calibration pool should reflect the behavior one aims to preserve. Future work can further test less aligned or domain-shifted candidate pools to study how broadly the same selection principles transfer; in deployment, target-specific gains should also be checked against broad-preservation metrics to ensure that the selected recipe remains appropriate beyond the primary target.

Quantizer and bit-width scope. Our main experiments focus on GPTQ-style calibration-data selection because GPTQ directly uses calibration strings to estimate activation statistics. We include AWQ and BNB-NF4 as quantizer-family checks, and the results suggest that some preservation signals transfer while score-based metrics remain quantizer-specific. A broader sweep over quantizers and bit-widths would further refine deployment-specific calibration recipes.

Interaction with post-hoc calibration. Our post-hoc analysis covers representative score-space methods to distinguish pre-quantization data selection from post-quantization score repair. These experiments show that the two stages can optimize different objectives and may be complementary. A fuller study of how to combine target-aware calibration selection with post-hoc calibration is a useful direction for future work.

Deployment risk. Uncertainty drift after quantization can affect downstream decisions in ways that are not visible from top-1 accuracy alone. In abstention, triage, educational QA, medical or legal assistance, and safety-filtering settings, shifted confidence or answerability boundaries may cause a system to answer when it should defer, or to suppress useful answers. This reinforces the need to audit quantized models not only for accuracy but also for the score behavior consumed by downstream decision modules.

Acknowledgments

This work was supported by the Start-up Grant of City University of Hong Kong (Dongguan). Generative AI tools were used only for language polishing and grammatical correction.

References

Jordan T. Ash, Chicheng Zhang, Akshay Krishnamurthy, John Langford, and Alekh Agarwal. 2020. Deep batch active learning by diverse, uncertain gradient lower bounds. In *International Conference on Learning Representations*.

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

Yonatan Bisk, Rowan Zellers, Ronan Le Bras, Jianfeng Gao, and Yejin Choi. 2020. PIQA: Reasoning about physical commonsense in natural language. In *Proceedings of the Thirty-Fourth AAAI Conference on Artificial Intelligence*, pages 7432–7439.

Glenn W. Brier. 1950. Verification of forecasts expressed in terms of probability. *Monthly Weather Review*, 78(1):1–3.

C. K. Chow. 1970. On optimum recognition error and reject tradeoff. *IEEE Transactions on Information Theory*, 16(1):41–46.

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, Volume 1 (Long and Short Papers)*, pages 2924–2936, Minneapolis, Minnesota. 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*.

Shrey Desai and Greg Durrett. 2020. Calibration of pre-trained transformers. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 295–302, Online. Association for Computational Linguistics.

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

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. QLoRA: Efficient finetuning of quantized LLMs. In *Advances in Neural Information Processing Systems*.

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 *International Conference on Learning Representations*.

Vage Egiazarian, Andrei Panferov, Denis Kuznedelev, Elias Frantar, Artem Babenko, and Dan Alistarh. 2024. Extreme compression of large language models via additive quantization. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 12284–12303. PMLR.

Ran El-Yaniv and Yair Wiener. 2010. On the foundations of noise-free selective classification. *Journal of Machine Learning Research*, 11:1605–1641.

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

Yonatan Geifman and Ran El-Yaniv. 2017. Selective classification for deep neural networks. In *Advances in Neural Information Processing Systems*.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, and 1 others. 2024. The Llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. 2017. On calibration of modern neural networks. In *Proceedings of the 34th International Conference on Machine Learning*, pages 1321–1330.

Liang He, Jingbo Wen, Qishi Zhan, Yixiong Chen, Kangning Cui, Qizhen Lan, and Xilu Wang. 2026. Budgetdraft: Acceptance-aware multi-view training for sparse-kv speculative decoding. *arXiv preprint arXiv:2606.00144*.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b. *arXiv preprint arXiv:2310.06825*.

Zhengbao Jiang, Frank F. Xu, Jun Araki, and Graham Neubig. 2020. How can we know what language models know? *Transactions of the Association for Computational Linguistics*, 8:423–438.

Saurav Kadavath, Tom Conerly, Amanda Askell, Tom Henighan, Dawn Drain, Ethan Perez, Nicholas Schiefer, Zac Hatfield-Dodds, Nova DasSarma, Eli Tran-Johnson, Scott Johnston, Sheer El-Showk, Andy Jones, Nelson Elhage, Tristan Hume, Anna Chen, Yuntao Bai, Sam Bowman, Stanislav Fort, and 17 others. 2022. Language models (mostly) know what they know. *arXiv preprint arXiv:2207.05221*.

Sanyam Kapoor, Nate Gruver, Manley Roberts, Katherine Collins, Arka Pal, Umang Bhatt, Adrian Weller, Samuel Dooley, Micah Goldblum, and Andrew G Wilson. 2024a. Large language models must be taught to know what they don't know. *Advances in Neural Information Processing Systems*, 37:85932–85972.

Sanyam Kapoor, Nate Gruver, Manley Roberts, Arka Pal, Samuel Dooley, Micah Goldblum, and Andrew Wilson. 2024b. Calibration-tuning: Teaching large language models to know what they don't know. In *Proceedings of the 1st Workshop on Uncertainty-Aware NLP (UncertaiNLP 2024)*, pages 1–14.

Sehoon Kim, Coleman Richard Charles Hooper, Amir Gholami, Zhen Dong, Xiuyu Li, Sheng Shen, Michael W. Mahoney, and Kurt Keutzer. 2024. SqueezeLLM: Dense-and-sparse quantization. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 23901–23923. PMLR.

Meelis Kull, Miquel Perello Nieto, Markus Kängsepp, Telmo Silva Filho, Hao Song, and Peter Flach. 2019. Beyond temperature scaling: Obtaining well-calibrated multiclass probabilities with dirichlet calibration. In *Advances in Neural Information Processing Systems*.

Ananya Kumar, Percy Liang, and Tengyu Ma. 2019. Verified uncertainty calibration. In *Advances in Neural Information Processing Systems*.

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

Stephanie Lin, Jacob Hilton, and Owain Evans. 2022. TruthfulQA: Measuring how models mimic human

falsehoods. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3214–3252, Dublin, Ireland. Association for Computational Linguistics.

Meta AI. 2024. Llama 3.2 model card. Model card for the Llama 3.2 text-only model collection.

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*, pages 2381–2391, Brussels, Belgium. Association for Computational Linguistics.

Matthias Minderer, Josip Djolonga, Rob Romijnders, Frances Ann Hubis, Xiaohua Zhai, Neil Houlsby, Dustin Tran, and Mario Lucic. 2021. Revisiting the calibration of modern neural networks. In *Advances in Neural Information Processing Systems*.

Mahdi Pakdaman Naeini, Gregory F. Cooper, and Milos Hauskrecht. 2015. Obtaining well calibrated probabilities using bayesian binning. In *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence*, pages 2901–2907.

Alexandru Niculescu-Mizil and Rich Caruana. 2005. Predicting good probabilities with supervised learning. In *Proceedings of the 22nd International Conference on Machine Learning*, pages 625–632.

Jeremy Nixon, Michael W. Dusenberry, Ghassen Jerfel, Timothy Nguyen, Jeremiah Liu, Linchuan Zhang, and Dustin Tran. 2019. Measuring calibration in deep learning. In *CVPR Workshops*.

Yaniv Ovadia, Emily Fertig, Jie Ren, Zachary Nado, D. Sculley, Sebastian Nowozin, Joshua V. Dillon, Balaji Lakshminarayanan, and Jasper Snoek. 2019. Can you trust your model's uncertainty? evaluating predictive uncertainty under dataset shift. In *Advances in Neural Information Processing Systems*.

John C. Platt. 1999. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. *Advances in Large Margin Classifiers*, pages 61–74.

Irina Proskurina, Luc Brun, Guillaume Metzler, and Julien Velcin. 2024. When quantization affects confidence of large language models? In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 1918–1928, Mexico City, Mexico. Association for Computational Linguistics.

Qwen Team. 2024. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*.

Pranav Rajpurkar, Robin Jia, and Percy Liang. 2018. Know what you don't know: Unanswerable questions for SQuAD. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 784–789, Melbourne, Australia. Association for Computational Linguistics.

Ozan Sener and Silvio Savarese. 2018. Active learning for convolutional neural networks: A core-set approach. In *International Conference on Learning Representations*.

Burr Settles. 2009. Active learning literature survey. Computer Sciences Technical Report 1648, University of Wisconsin–Madison.

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 *International Conference on Learning Representations*.

Zishan Shao, Lixun Zhang, Kangning Cui, Yixiao Wang, Ting Jiang, Hancheng Ye, Qinsi Wang, Zhixu Du, Yuzhe Fu, Fan Yang, and 1 others. 2026. Decodeshare: Tracing the shared subspace of llm decode-time decisions. In *Proceedings of the International Conference on Machine Learning*.

Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. 2019. CommonsenseQA: A question answering challenge targeting commonsense knowledge. 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)*, pages 4149–4158, Minneapolis, Minnesota. Association for Computational Linguistics.

Albert Tseng, Jerry Chee, Qingyao Sun, Volodymyr Kuleshov, and Christopher De Sa. 2024. QuIP#: Even better LLM quantization with hadamard incoherence and lattice codebooks. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 48630–48656. PMLR.

Miles Williams and Nikolaos Aletras. 2024. On the impact of calibration data in post-training quantization and pruning. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10100–10118, Bangkok, Thailand. Association for Computational Linguistics.

Miles Williams, George Chrysostomou, and Nikolaos Aletras. 2025. Self-calibration for language model quantization and pruning. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 10149–10167.

Wenhao Wu, Zishan Shao, Kangning Cui, Jinhee Kim, Yixiao Wang, Hancheng Ye, Danyang Zhuo, and Yiran Chen. 2026. Flashsvd v1. 5: Making low-rank transformers inference actually fast. *arXiv preprint arXiv:2605.08314*.

Guangxuan Xiao, Ji Lin, Mickael Seznec, Hao Wu, Julien Demouth, and Song Han. 2023. SmoothQuant: Accurate and efficient post-training quantization for

large language models. In *Proceedings of the 40th International Conference on Machine Learning*.

Yuxin Xiao, Paul Pu Liang, Umang Bhatt, Willie Neiswanger, Ruslan Salakhutdinov, and Louis-Philippe Morency. 2022. Uncertainty quantification with pre-trained language models: A large-scale empirical analysis. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 7273–7284, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Zhewei Yao, Reza Yazdani Aminabadi, Minjia Zhang, Xiaoxia Wu, Conglong Li, and Yuxiong He. 2022. ZeroQuant: Efficient and affordable post-training quantization for large-scale transformers. In *Advances in Neural Information Processing Systems*, volume 35.

Bianca Zadrozny and Charles Elkan. 2002. Transforming classifier scores into accurate multiclass probability estimates. In *Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 694–699.

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 Annual Meeting of the Association for Computational Linguistics*, pages 4791–4800, Florence, Italy. Association for Computational Linguistics.

Tony Z. Zhao, Eric Wallace, Shi Feng, Dan Klein, and Sameer Singh. 2021. Calibrate before use: Improving few-shot performance of language models. In *Proceedings of the 38th International Conference on Machine Learning*, pages 12697–12706.

Mingyu Zhong, Guanchu Wang, Yu-Neng Chuang, and Na Zou. 2025. Quantized can still be calibrated: A unified framework to calibration in quantized large language models. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 30503–30517, Vienna, Austria. Association for Computational Linguistics.

A Additional Experimental Results

This appendix collects evidence that supports the main text but is too detailed to include there. It is organized as follows: §A.1 reports stability checks and ablations on the answerability-boundary target; §A.2 reports broad-preservation diagnostics; §A.3 analyzes post-hoc calibration; and §A.4 presents the AWQ quantizer-family extension. Throughout, §A.1 uses the answerability-boundary ranking, while §A.2 uses broad FP-behavior preservation. A method can therefore be strong under one target and only average under another. We keep the most representative evidence in this main appendix, while additional diagnostics for broad preservation, post-hoc calibration, and AWQ transfer are collected in Appendix B.

A.1 Stability and Ablations on the Answerability-Boundary Target

Stability of the core result. Table 6 repeats the SQuAD2 answerability analysis after removing the Llama-3.2-1B stress case. The ordering remains consistent: DPQ-s128-r75 stays first, and its FP agreement rises to 0.9360.

Answerability distribution diagnostics. Figure 5 provides the same SQuAD2 answerability-distribution diagnostic for Qwen2.5-7B and Mistral-7B-v0.3, comparing full precision, GPTQ-WikiText, and DPQ-r75 via the ECDF of $p(\text{answerable})$. Together with Figure 3, these plots show that the effect is not a single-model artifact: DPQ-r75 generally reduces quantization-induced answerability-distribution drift, while the direction and magnitude of the residual drift are model-dependent.

Ablations on the boundary target. Table 7 reports the full ablation grid grouped by axis. The mixture ratio is non-monotonic: both r0 and r100 trail r75, indicating that boundary mass should be high but not exclusive. The size sweep shows that composition of the 128 examples matters more than scaling to 64 or 256. Boundary-only and boundary-random fall below r75, supporting the role of generic anchors; each single-signal component is informative but none matches the mixed recipe. Negative controls (HighNLL-QA, LowDoubt-QA) trail r75 by ≥ 0.09 on boundary-accuracy deviation, distinguishing difficulty from doubt (Appendix C.4). Strong data-selection baselines—Activation k-center, Uncertainty-only, TaskRandom, RandomQA, and self-calibration—improve over generic WikiText/C4 calibration but remain below DPQ-r75 on the boundary-specific aggregate.

A.2 Broad-Preservation Diagnostics

When the target shifts from SQuAD2 answerability boundaries to broad FP-behavior preservation, the boundary-heavy DPQ-r75 mixture is not expected to dominate. Table 8 expands Table 4 on the six answerable extra-MCQA datasets and is the most representative broad-target leaderboard supporting the target-dependent claim of Section 5.4. Confidence-only, TaskRandom, HighNLL-QA, entropy-only, and uncertainty-only all sit above

Method	Rank ↓	Agr. ↑	AccΔ ↓	RateΔ ↓	JSD ↓
DPQ-s128-r75	6.26	0.9360	0.0331	0.0094	0.0088
DPQ-answerability-only	7.50	0.9250	0.0569	0.0267	0.0100
DPQ-entropy-only	7.66	0.9263	0.0891	0.0446	0.0091
DPQ-boundary-random	8.26	0.9326	0.0411	0.0183	0.0093
DPQ-boundary-only	8.61	0.9316	0.0460	0.0210	0.0099
GPTQ-ActivationKCenter	9.39	0.9294	0.0606	0.0291	0.0105
Uncertainty-only	9.67	0.9139	0.1197	0.0590	0.0109
DPQ-s64-r50	9.71	0.9224	0.0769	0.0364	0.0100

Table 6: SQuAD2 answerability-boundary preservation after removing the Llama-3.2-1B stress case. The main r75 result remains stable.

Group	Method	Rank ↓	Agr. ↑	AccΔ ↓	RateΔ ↓	JSD ↓
Reference	DPQ-s128-r75	6.20	0.8495	0.2125	0.1000	0.0158
Ratio / size	DPQ-s64-r50	10.18	0.8245	0.2800	0.1382	0.0197
	DPQ-s256-r50	11.18	0.8240	0.2870	0.1425	0.0235
	DPQ-s128-r0	10.31	0.8097	0.3205	0.1598	0.0220
	DPQ-s128-r25	12.06	0.8106	0.3277	0.1639	0.0214
	DPQ-s128-r50	13.12	0.8140	0.3210	0.1602	0.0258
	DPQ-s128-r100	12.44	0.8080	0.3340	0.1655	0.0215
Boundary mix	DPQ-boundary-random	8.15	0.8413	0.2310	0.1135	0.0174
	DPQ-boundary-only	9.06	0.8335	0.2520	0.1242	0.0197
Single signal	DPQ-answerability-only	8.18	0.8286	0.2592	0.1281	0.0205
	DPQ-entropy-only	8.68	0.8270	0.2940	0.1470	0.0227
	DPQ-confidence-only	11.30	0.8033	0.3460	0.1725	0.0221
Neg. control	HighNLL-QA	11.11	0.8191	0.3047	0.1514	0.0201
	LowDoubt-QA	12.09	0.8177	0.3110	0.1532	0.0202
Data selection	GPTQ-ActivationKCenter	9.39	0.8360	0.2530	0.1255	0.0187
	Uncertainty-only	9.36	0.8261	0.2937	0.1461	0.0181
	GPTQ-TaskRandom	10.81	0.8240	0.2920	0.1445	0.0216
	GPTQ-RandomQA	11.09	0.8266	0.2733	0.1351	0.0231
	GPTQ-SelfCalib	13.24	0.8129	0.3037	0.1509	0.0268
	GPTQ-C4	15.78	0.7594	0.4107	0.2034	0.0342
	GPTQ-WikiText	17.29	0.7446	0.4553	0.2271	0.0387

Table 7: Boundary-target ablation grid on SQuAD2 answerability. Methods are grouped by ablation axis; columns follow Table 2. AccΔ and RateΔ denote boundary-accuracy and answerability-rate deviation, respectively.

DPQ-r75; BNB-NF4 has the strongest FP agreement and JSD but is not the best average rank, showing that preserving one metric family is insufficient under the full multi-metric objective. Additional diagnostics on calibration-gap prevalence and correct-vs-wrong confidence decomposition are reported in Appendix B.1; they further show that quantization can change confidence and margin behavior even when top-1 accuracy gives an incomplete picture.

A.3 Post-Hoc Calibration Analysis

Post-hoc on both suites. Table 9 extends the main post-hoc analysis (Table 5) to both the old-core and extra-MCQA suites, providing the most representative summary of how each post-hoc family behaves on FP-behavior metrics. Adaptive temperature leaves accuracy and FP agreement unchanged but increases JSD and margin drift, consistent with Proposition 5. Flexible calibrators (vector, matrix, Dirichlet, isotonic) raise accuracy — particularly on old-core — while reducing FP agreement and increasing distributional drift, consistent with Proposition 6. The two stages thus serve different objectives: post-hoc calibration reshapes scores, whereas pre-quantization data selection controls which behavior is preserved in the first place. Appendix B.2 reports an optimally fitted temperature-scaling check, which supports the same complementary-not-substitute conclusion.

A.4 AWQ Quantizer-Family Extension

The AWQ extension covers all eight models, three old-core datasets, and four AWQ calibration

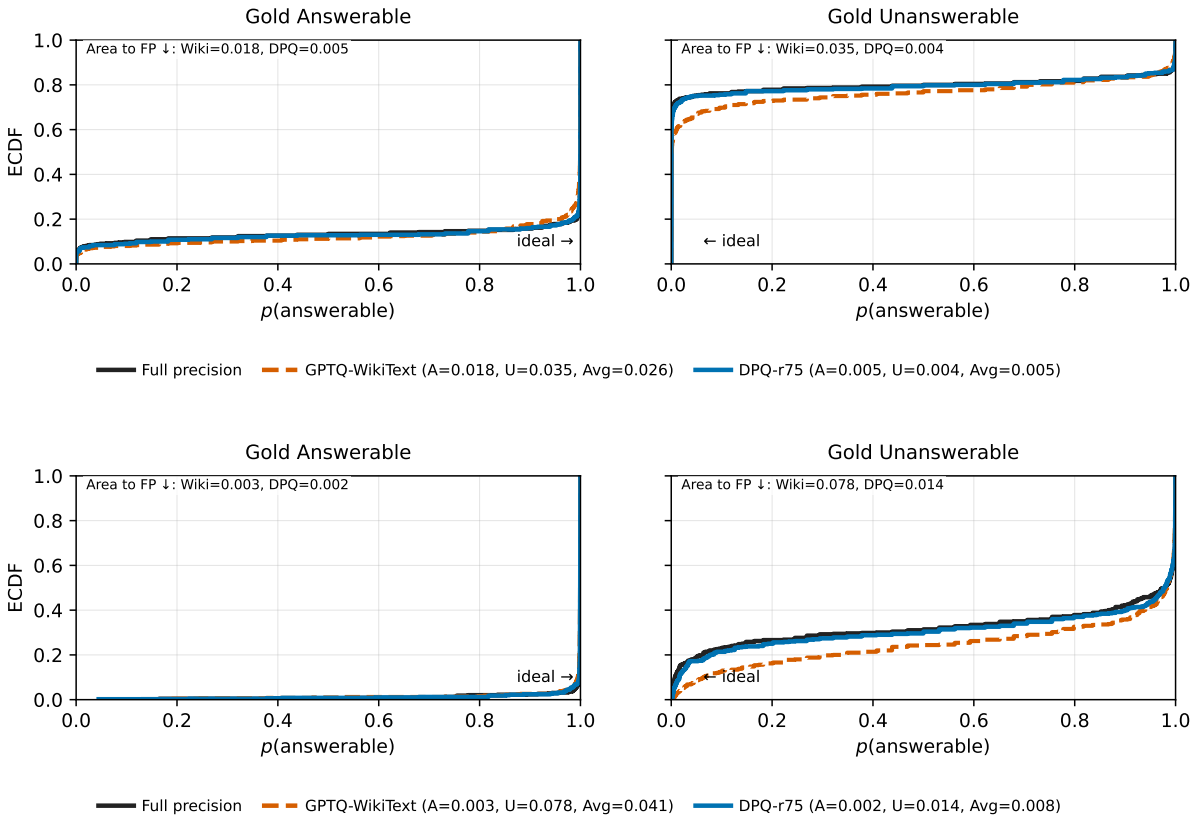

Figure 5: Additional SQuAD2 answerability ECDFs for Qwen2.5-7B (top) and Mistral-7B-v0.3 (bottom). Lower area-to-FP indicates closer preservation of FP answerability behavior.

recipes (WikiText, RandomQA, DPQ-r75, DPQ-r50), totaling $8 \times 3 \times 4 = 96$ evaluations, to test whether the calibration-data signal observed under GPTQ transfers to a different quantizer family. On AWQ, score-based metrics (Acc, ECE, NLL, Brier) do not always favor the same recipe as FP-behavior metrics (FP agreement, JSD-to-FP), so we report aggregate metric averages and rank-based summaries separately in Appendix B.3. Aggregate summaries in Appendix B.3 show the same mixed pattern: generic calibration can win score-based metrics, while DPQ variants are stronger on several FP-behavior and rank-based summaries. On the 7B/8B subset (Qwen2.5-7B, Llama-3.1-8B, Mistral-7B-v0.3), DPQ-r75 additionally leads on accuracy, Brier, FP agreement, and the rank-based summaries. The most informative single view is the per-dataset breakdown below.

AWQ per-dataset breakdown. Table 10 reports the full 3-dataset $\times$ 4-recipe AWQ grid. The transfer to AWQ is dataset-dependent. On ARC-Challenge, DPQ-r50 has the best accuracy and FP agreement and DPQ-r75 the best Brier score, so DPQ variants dominate. On SQuAD2 answerability the picture inverts: WikiText takes the best accuracy, ECE, NLL, and Brier, while DPQ variants (r50 marginally above r75) retain only the FP-agreement lead. On TruthfulQA, WikiText is strongest on every metric reported here. This pattern is consistent with the main message: the calibration-data signal is real under AWQ, but the optimal recipe is quantizer- and target-dependent rather than fixed at r75.

B Additional Diagnostic Analyses

This appendix collects additional diagnostic analyses that further decompose or stress-test the results summarized in Appendix A. These checks provide secondary decompositions and stress tests, while Appendix A keeps the primary evidence focused on the main target-specific claims.

B.1 Additional Broad-Preservation Diagnostics

This subsection provides two compact diagnostics supporting the claim that quantization changes score behavior in ways not captured by top-1 accuracy alone. Table 11 extends the main degradation

Method	Family	Rank ↓	Top-5 ↑	Agr. ↑	JSD ↓	MarginΔ ↓
DPQ-confidence-only	DPQ component	8.96	7	0.8357	0.0284	0.1447
GPTQ-TaskRandom	Selection baseline	9.12	11	0.8361	0.0276	0.1452
HighNLL-QA	Negative control	9.18	6	0.8327	0.0294	0.1478
DPQ-entropy-only	DPQ component	9.30	10	0.8324	0.0296	0.1473
Uncertainty-only	Selection baseline	9.31	4	0.8318	0.0297	0.1471
LowDoubt-QA	Negative control	9.47	9	0.8328	0.0286	0.1472
GPTQ-ActivationKCenter	Selection baseline	9.53	12	0.8349	0.0299	0.1483
DPQ-s128-r50	DPQ mixed	9.92	4	0.8322	0.0293	0.1454
BNB-NF4	Quantizer baseline	10.53	12	0.8637	0.0241	0.1320
DPQ-boundary-random	DPQ mixed	10.66	8	0.8279	0.0311	0.1513
DPQ-s128-r75	DPQ mixed	11.88	3	0.8180	0.0324	0.1562

Table 8: Expanded broad-MCQA preservation on the six answerable datasets. The table adds method family and margin drift to the main broad-MCQA summary in Table 4.

Suite	Method	Rows	Acc. ↑	Agr. ↑	JSD ↓	MarginΔ ↓
Old-core	Base	72	0.6196	0.7906	0.0364	0.1609
	Adaptive temp.	72	0.6196	0.7906	0.0605	0.2892
	Option bias	72	0.5991	0.7811	0.0410	0.1876
	Vector	72	0.7985	0.6875	0.1360	0.3099
	Matrix	72	0.7925	0.6796	0.1404	0.3146
	Dirichlet	72	0.7929	0.6817	0.1403	0.3162
	Isotonic	72	0.7983	0.6883	0.1398	0.2990
Extra-MCQA	Base	1056	0.6833	0.8245	0.0320	0.1533
	Adaptive temp.	1056	0.6833	0.8245	0.0445	0.2385
	Option bias	1056	0.6985	0.8147	0.0328	0.1591
	Vector	1056	0.7029	0.7773	0.0519	0.2344
	Matrix	1056	0.6992	0.7723	0.0550	0.2380
	Dirichlet	1056	0.6989	0.7740	0.0548	0.2400
	Isotonic	1056	0.6991	0.7710	0.0561	0.2329

Table 9: Post-hoc family summary on old-core and extra-MCQA suites. Flexible calibrators can improve accuracy but often reduce agreement with FP and increase distributional drift.

summary with confidence shift and margin drift. Table 12 further decomposes confidence on correct and wrong predictions, showing that average calibration metrics can hide changes in the separation between reliable and unreliable outputs.

Calibration-gap prevalence. Table 11 extends Table 1 with confidence shift and margin drift. Quantization frequently worsens not only accuracy but also calibration-sensitive and drift metrics. The last two columns highlight that generic-text calibration (WikiText, C4) induces the largest mean confidence and margin shifts, and no row eliminates degradation across all columns.

Correct-vs-wrong confidence decomposition. Table 12 splits confidence into a gap on correct predictions and overconfidence on wrong ones. Quantized methods often reduce confidence on correct predictions while keeping wrong-prediction confidence high on extra-MCQA, so a method may appear only mildly shifted in average confidence while degrading the separation between reliable and unreliable outputs.

B.2 Additional Post-Hoc Calibration Checks

This subsection adds an optimally fitted temperature-scaling check to complement Table 9. The goal is to test whether the temperature result in the main post-hoc analysis is simply due to an under-tuned scalar temperature.

Optimally fitted temperature. Table 13 fits one optimal temperature per model–dataset setting. Temperature scaling improves ECE and NLL deltas relative to FP, but it leaves accuracy, FP agreement, and boundary decisions unchanged by construction. This supports the view that temperature scaling can polish proper scores but cannot repair decision-surface drift.

B.3 AWQ Aggregate Summaries

Tables 14 and 15 report aggregate AWQ metric averages and rank-based summaries, complementing the per-dataset breakdown in Table 10. Across

Dataset	AWQ recipe	n	Acc. ↑	ECE ↓	NLL ↓	Brier ↓	Agr. ↑
ARC-Challenge	WikiText	8	0.6994	0.1185	0.8645	0.4133	0.8516
	RandomQA	8	0.7082	**0.0920**	**0.8191**	0.3974	0.8637
	DPQ-r75	8	0.7153	0.1040	0.8360	**0.3952**	0.8662
	DPQ-r50	8	**0.7174**	0.1136	0.8373	0.4008	**0.8792**
SQuAD2 answerability	WikiText	8	**0.6474**	**0.2122**	**1.0830**	**0.5330**	0.8483
	RandomQA	8	0.6429	0.2237	1.2086	0.5472	0.8398
	DPQ-r75	8	0.6301	0.2501	1.2885	0.5687	0.8545
	DPQ-r50	8	0.6324	0.2502	1.3269	0.5830	**0.8558**
TruthfulQA	WikiText	8	**0.5332**	**0.2366**	**1.7355**	**0.6887**	**0.8040**
	RandomQA	8	0.5300	0.2460	1.7858	0.6985	0.7945
	DPQ-r75	8	0.4991	0.2513	1.7787	0.7111	0.7930
	DPQ-r50	8	0.4841	0.2681	1.8821	0.7442	0.7971

Table 10: AWQ per-dataset × recipe breakdown averaged over the 8 models. Best values for each dataset and metric are bolded. Agr. denotes agreement with the full-precision model. The preferred recipe varies by dataset and metric, showing that AWQ transfer is target- and metric-dependent rather than fixed to a single calibration recipe.

Method	n	Worse than FP (%, ↓)				Mean drift (↓)	
		Acc.	ECE	NLL	Brier	Conf. shift	MarginΔ
BNB-NF4	72	84.7	66.7	84.7	81.9	0.0190	0.1376
GPTQ-RandomQA	72	83.3	61.1	76.4	86.1	0.0373	0.1500
GPTQ-WikiText	72	76.4	58.3	63.9	83.3	0.0583	0.1901
GPTQ-C4	72	83.3	65.3	68.1	77.8	0.0493	0.1812
DPQ-s128-r75	72	80.6	62.5	70.8	83.3	0.0363	0.1543
DPQ-s128-r50	72	79.2	58.3	65.3	80.6	0.0321	0.1474
DPQ-confidence-only	72	80.6	62.5	66.7	81.9	0.0323	0.1468
DPQ-entropy-only	72	76.4	73.6	75.0	86.1	0.0322	0.1445
Uncertainty-only	72	77.8	59.7	73.6	80.6	0.0331	0.1473
GPTQ-TaskRandom	72	84.7	65.3	73.6	83.3	0.0320	0.1448

Table 11: Calibration-gap prevalence on all nine datasets. The first four metric columns follow Table 1; the last two columns report mean confidence shift and margin drift relative to FP.

all 8 models, WikiText or RandomQA win several score-based metrics while DPQ variants are stronger on FP-behavior metrics; the rank aggregate gives DPQ-r75 the best mean rank and top-2 frequency. The 7B/8B subset is cleaner: DPQ-r75 leads on accuracy, Brier, FP agreement, mean rank, and top-1/top-2 frequencies, with RandomQA only marginally better on ECE and JSD-to-FP. These aggregate results are consistent with stronger transfer on the 7B/8B-scale models, while smaller models introduce more variability.

C Formal Analysis and Proofs

This section gives a more complete formal account of the mechanism. Recall the goal is to make precise three narrower claims: low-margin examples are fragile under quantization; calibration distributions should match the uncertainty target; and post-hoc calibration is not equivalent to pre-quantization preservation.

C.1 Notation

For an input x with K candidate options, let $s_0(x) \in \mathbb{R}^K$ be the FP option-score vector and $p_0(x) = \mathrm{softmax}(s_0(x))$. For a quantized model calibrated with set D, write $s_D(x) = s_0(x) + \Delta_D(x)$ and $p_D(x) = \mathrm{softmax}(s_D(x))$. Let $i^*(x), j^*(x)$ denote the FP top option and runner-up. The logit margin is $\gamma(x) = z_{0,i^*}(x) - z_{0,j^*}(x)$, and the probability margin is $m(x) = p_{0,(1)}(x) - p_{0,(2)}(x)$.

C.2 Boundary Fragility

Proposition 3 (Top-two boundary flip condition)**.** *With logit margin $\gamma(x) > 0$, the relative ordering of i^* and j^* flips after quantization iff $\Delta_D(x, j^*) - \Delta_D(x, i^*) > \gamma(x)$. If $\|\Delta_D(x)\|_\infty < \gamma(x)/2$, then i^* remains above every other option and the top-1* FP *decision is preserved.*

Proof. Write $\delta_k = \Delta_D(x, k)$ for option index k.

Suite	Method	Acc.	Conf. correct	Correct gap	Conf. wrong
Old-core	FP	0.6483	0.8358	0.1642	0.7225
	DPQ-s128-r75	0.6136	0.8054	0.1946	0.6953
	DPQ-confidence-only	0.6338	0.8106	0.1894	0.6996
	DPQ-entropy-only	0.6418	0.8153	0.1847	0.7071
	GPTQ-RandomQA	0.6316	0.8089	0.1911	0.7042
	GPTQ-TaskRandom	0.6296	0.8087	0.1913	0.6999
	BNB-NF4	0.6239	0.8353	0.1647	0.7233
Extra-MCQA	FP	0.7232	0.8424	0.1576	0.7070
	DPQ-s128-r75	0.6892	0.8134	0.1866	0.6929
	DPQ-s128-r50	0.6898	0.8166	0.1834	0.6929
	DPQ-confidence-only	0.6893	0.8192	0.1808	0.6946
	DPQ-entropy-only	0.6861	0.8166	0.1834	0.6979
	GPTQ-RandomQA	0.6884	0.8063	0.1937	0.6817
	GPTQ-TaskRandom	0.6861	0.8116	0.1884	0.6884
	BNB-NF4	0.6945	0.8348	0.1652	0.7102

Table 12: Correct-vs-wrong confidence decomposition. "Correct gap" is $1 - \mathbb{E}[\text{conf} \mid \text{correct}]$; "Conf. wrong" is $\mathbb{E}[\text{conf} \mid \text{wrong}]$, which directly measures overconfidence on errors.

Suite	Method	n	ECE Δ pre→post	NLL Δ pre→post
Old-core	GPTQ-WikiText	24	$-0.009 \to -0.091$	$-0.050 \to -0.308$
	GPTQ-RandomQA	24	$-0.003 \to -0.097$	$+0.013 \to -0.314$
	DPQ-r75	24	$+0.002 \to -0.102$	$+0.054 \to -0.296$
	BNB-NF4	24	$+0.012 \to -0.091$	$+0.099 \to -0.313$
Extra-MCQA	GPTQ-WikiText	48	$+0.015 \to -0.048$	$+0.091 \to -0.037$
	GPTQ-RandomQA	48	$+0.002 \to -0.055$	$+0.046 \to -0.066$
	DPQ-r75	48	$+0.004 \to -0.056$	$+0.045 \to -0.071$
	BNB-NF4	48	$+0.015 \to -0.050$	$+0.078 \to -0.076$

Table 13: Effect of fitting one optimal temperature per model–dataset setting. Temperature improves proper-score deltas but does not change accuracy, FP agreement, or boundary decisions.

For the top-two pair, $z_D(x, i^*) - z_D(x, j^*) = \gamma(x) + \delta_{i^*} - \delta_{j^*}$, which is negative iff $\delta_{j^*} - \delta_{i^*} > \gamma(x)$. For any $k \neq i^*$, $z_0(x, i^*) - z_0(x, k) \geq \gamma(x)$, so if $\|\Delta_D(x)\|_\infty < \gamma(x)/2$ then $z_D(x, i^*) - z_D(x, k) > 0$. □

Corollary 1 (Flip mass is controlled by the margin distribution)**.** *Assume $\|\Delta_D(x)\|_\infty \leq \eta$ for all x in a test distribution T. Then* $\Pr_{x\sim T}[\arg\max z_D(x) \neq \arg\max z_0(x)] \leq \Pr_{x\sim T}[\gamma(x) \leq 2\eta]$.

C.3 Calibration Distribution and Target Mismatch

Let q denote the distribution over calibration strings used by the quantizer, and T the target evaluation distribution.

Proposition 4 (Target mismatch bound)**.** *For any bounded loss $\ell_D \in [0, 1]$, $|R_T(D) - R_q(D)| \leq \mathrm{TV}(T, q)$.*

Proof. By the variational characterization, $\sup_{0\leq f\leq 1} |\mathbb{E}_T f - \mathbb{E}_q f| = \mathrm{TV}(T, q)$. Taking $f = \ell_D$ gives the result. □

Mixture interpretation. Write $q_r = r\, q_{\text{bdry}} + (1 - r)\, q_{\text{anchor}}$ and $T_\theta = \theta q_{\text{bdry}} + (1 - \theta) q_{\text{anchor}}$. If q_{bdry} and q_{anchor} have disjoint support, $\mathrm{TV}(T_\theta, q_r) = |\theta - r|$. This is the form used in Proposition 2 of the main text.

Finite-sample coverage. Calibration sets are empirical samples. If random QA calibration samples n examples from a distribution with boundary mass $\rho = \Pr[B]$, then $\Pr[N_B = 0] = (1 - \rho)^n \leq \exp(-n\rho)$, so $n \geq \log(1/\delta)/\rho$ is needed to see a boundary example with probability $\geq 1 - \delta$. When ρ is small, random calibration needs many samples to cover the fragile region. DPQ bypasses this by selecting boundary examples directly using the FP model.

C.4 Boundary vs Hard-Example Mining

High NLL does not imply low margin: a binary example with FP probabilities $(1 - \epsilon, \epsilon)$ and gold label option 2 has NLL $-\log \epsilon$ (arbitrarily large) but margin $1 - 2\epsilon$ (close to 1, far from the boundary). Conversely, an example with FP probabilities $(1/2 + \epsilon, 1/2 - \epsilon)$ and gold label option 1 has moderate NLL but margin 2ϵ (arbitrarily small). Hard-example mining on gold-label NLL therefore selects a subset distinct from boundary mining; an example can be confidently wrong (high NLL, large margin) without being uncertain.

C.5 Why Post-hoc Calibration Is Not Equivalent

Proposition 5 (Temperature scaling preserves the decision surface)**.** *For any $z \in \mathbb{R}^K$ and $T > 0$,* $\arg\max_i \mathrm{softmax}(z/T)_i = \arg\max_i z_i$. *There-*

Scope	AWQ method	Acc. ↑	ECE ↓	NLL ↓	Brier ↓	Agr. ↑	JSD ↓
All 8 models	WikiText	0.6267	0.1891	**1.2277**	**0.5450**	0.8346	0.0325
	RandomQA	**0.6270**	**0.1872**	1.2712	0.5477	0.8327	0.0336
	DPQ-r75	0.6148	0.2018	1.3011	0.5583	0.8379	**0.0320**
	DPQ-r50	0.6113	0.2106	1.3488	0.5760	**0.8440**	0.0332
7B/8B subset	WikiText	0.6987	0.1857	**1.2220**	0.4784	0.8950	0.0262
	RandomQA	0.7105	**0.1795**	1.2368	0.4618	0.9116	**0.0210**
	DPQ-r75	**0.7160**	0.1802	1.2386	**0.4558**	**0.9138**	0.0213
	DPQ-r50	0.7083	0.1842	1.2590	0.4681	0.9034	0.0224

Table 14: AWQ metric averages over old-core datasets, split by model scope. The 7B/8B subset is Qwen2.5-7B, Llama-3.1-8B, and Mistral-7B-v0.3. Agr. denotes agreement with the full-precision model.

Scope	AWQ method	Mean rank ↓	Top-1 freq. ↑	Top-2 freq. ↑
All 8 models	DPQ-r75	**2.25**	**29.2%**	**66.7%**
	RandomQA	2.41	8.3%	62.5%
	WikiText	2.66	16.7%	45.8%
	DPQ-r50	2.69	4.2%	50.0%
7B/8B subset	DPQ-r75	**1.97**	**42.9%**	**73.0%**
	RandomQA	2.43	27.0%	47.6%
	WikiText	2.76	19.1%	46.0%
	DPQ-r50	2.79	11.1%	36.5%

Table 15: Rank-based AWQ summary by model scope.

fore, scalar temperature scaling cannot repair a top-1 or answerability-boundary flip in the quantized model.

Proposition 6 (Accuracy repair and FP preservation can conflict)**.** *There exist binary examples for which a post-hoc mapping improves accuracy relative to a quantized model while decreasing agreement with the* FP *model.*

Proof. Take $p_0 = (0.55, 0.45)$, $p_Q = (0.45, 0.55)$, gold label 2. A post-hoc map producing $p_H = (0.10, 0.90)$ lowers NLL on the gold label but has larger JSD from p_0. A map restoring agreement with p_0 would reduce accuracy on this example. Improving accuracy and preserving the FP decision surface conflict here. □

C.6 Correct-vs-Wrong Decomposition

Let A be the event that the prediction is correct, $c(x) = \max_i p_i(x)$, and $a = \Pr[A]$. With $G_{\text{corr}} = \mathbb{E}[1-c \mid A]$ and $O_{\text{wrong}} = \mathbb{E}[c \mid A^c]$, $\mathbb{E}[c] = a(1-G_{\text{corr}}) + (1-a)O_{\text{wrong}}$. Two methods with similar ECE may differ on which component dominates.

D Reproducibility Details

Experimental inventory and completeness. Table 16 lists the experimental groups. The final merged result files contain zero missing entries for 552 old-core pre-quantization rows and 1104 extra-MCQA pre-quantization rows.

Model sizes and checkpoints. Table 17 reports the model families, nominal parameter scales, and public checkpoints used in the experiments. All models are instruction-tuned language models in the 0.5B–8B range. We report nominal model scales because the exact trainable parameter count can depend on checkpoint metadata and tokenizer/configuration conventions, while the deployment-relevant scale is the advertised checkpoint size.

Compute infrastructure and budget. All experiments were conducted on a single NVIDIA GeForce RTX 5090 GPU with 32GB memory. The study uses post-training quantization, calibration-data selection, inference-based evaluation, score-space post-hoc analysis, and AWQ extension experiments; no model was trained from scratch or instruction-tuned as part of the main method. The dominant costs are repeated 4-bit GPTQ/AWQ quantization and deterministic option-scoring evaluation over saved datasets. We retained all prediction JSONL files, summary CSVs, merged reports, and derived-analysis tables, but cleaned quantized checkpoint directories and Hugging Face caches during experimentation. We report the hardware environment rather than a precise GPU-hour total because exploratory runs, sanity checks, and reruns were not logged with a separate compute ledger; the final experiments are reproducible on the single-GPU infrastructure described above.

Dataset sizes and splits. Table 18 reports the deterministic evaluation sizes used after filtering and subsampling. Calibration examples are drawn from training or calibration pools only; evaluation examples are never used to construct DPQ calibration strings.

Implementation details. The released scripts use teacher-forced option scoring. For each op-

Group	Purpose	Coverage
Full precision	unquantized reference	all 9 datasets
BNB-NF4	4-bit NF4 quantizer baseline	all 9 datasets
GPTQ-WikiText / GPTQ-C4	generic text calibration	WikiText all 9; C4 extended
GPTQ-RandomQA / TaskRandom	task-formatted random QA	all 9 datasets
DPQ-r0/r25/r50/r75/r100	high-doubt ratio sweep at s=128	old-core + extra-MCQA
DPQ-s64/s128/s256-r50	calibration-size sweep	old-core + extra-MCQA
boundary-only / boundary-random	boundary component controls	old-core + extra-MCQA
confidence / entropy / answerability-only	single-signal DPQ components	old-core + extra-MCQA
low-doubt / high-NLL QA	negative controls	old-core + extra-MCQA
activation k-center / self-calib	strong data-selection baselines	old-core + extra-MCQA
temperature / score-space post-hoc	post-hoc controls	selected core controls
AWQ extension	quantizer-family check	old core, 8 models, 4 recipes

Table 16: Inventory of experimental groups.

Tag	Nominal scale	Checkpoint
`qwen25_05b`	0.5B	`Qwen/Qwen2.5-0.5B-Instruct`
`qwen25_1p5b`	1.5B	`Qwen/Qwen2.5-1.5B-Instruct`
`qwen25_3b`	3B	`Qwen/Qwen2.5-3B-Instruct`
`qwen25_7b`	7B	`Qwen/Qwen2.5-7B-Instruct`
`llama32_1b_instruct`	1B	`meta-llama/Llama-3.2-1B-Instruct`
`llama32_3b_instruct`	3B	`meta-llama/Llama-3.2-3B-Instruct`
`llama31_8b_instruct`	8B	`meta-llama/Llama-3.1-8B-Instruct`
`mistral7b_v03`	7B	`mistralai/Mistral-7B-Instruct-v0.3`

Table 17: Model identifiers and nominal parameter scales.

Dataset	Suite	Eval. examples
ARC-Challenge	Old-core	299
SQuAD2 answerability	Old-core	1000
TruthfulQA	Old-core	817
ARC-Easy	Extra-MCQA	570
BoolQ	Extra-MCQA	1000
PIQA	Extra-MCQA	1000
HellaSwag	Extra-MCQA	1000
OpenBookQA	Extra-MCQA	500
CommonsenseQA	Extra-MCQA	1000

Table 18: Evaluation sizes after deterministic filtering and subsampling.

tion, the option text is appended to the prompt; only option tokens are scored, and the option score is the mean token log-probability unless explicitly disabled. All evaluation runs use left padding, maximum sequence length 2048, and deterministic JSONL prediction files containing scores, probabilities, predicted option, gold option, confidence, entropy, and an unknown flag for SQuAD2 answerability.

GPTQ and BNB settings. GPTQ runs use 4-bit quantization through the Transformers `GPTQConfig` interface with model sequence length 512, batch size 1, and at most 128 calibration examples unless the size ablation specifies 64 or 256. The `desc_act` flag is off by default. The BNB baseline uses 4-bit NF4 with double quantization and the model dtype as compute dtype (bfloat16 if supported, else float16).

DPQ calibration construction. Candidate boundary pools are built from ARC-Challenge training examples and SQuAD2 training examples, with 512 candidates from each source before model-specific scoring; SQuAD2 candidates are balanced between answerable and unanswerable cases. Formal calibration files are built with seed 4242, and the candidate-pool construction uses seed 2027. DPQ ranks candidates by margin/confidence/entropy/answerability/NLL depending on the variant. For boundary strings, the selected prompt is concatenated with the FP top options. Mixed recipes allocate a fraction of the calibration set to boundary/high-doubt strings and the remainder to WikiText/random-QA anchors. This formatting is kept fixed across DPQ variants, so the reported ratio, size, component, and negative-control ablations compare selection choices under a shared boundary-string construction; fully separating formatting effects from selection effects is a natural follow-up control.

Prompt format and option scoring. Multiple-choice prompts are formatted as a question followed by labeled choices and the string "The correct answer is". Candidate options are scored as label continuations. For SQuAD2 answerability, each prompt contains the passage, question, and two options: "Answerable from the passage" and "Unanswerable from the passage". Scores are length-normalized by default and converted to option probabilities with a softmax.

Post-hoc controls. Temperature and score-space post-hoc methods are evaluated from saved logits/probabilities. Temperature scaling is fitted per evaluation setting for the additional temperature check. Flexible score-space calibrators, including option-bias, vector, matrix, Dirichlet, and isotonic mappings, are evaluated on saved option scores.

Artifacts, licenses, and retained outputs. We use public model checkpoints, quantization implementations, and standard NLP benchmarks for research evaluation only, without redistributing the original checkpoints or datasets. We retained prediction JSONL files, summary CSVs, merged reports, and derived-analysis tables; quantized checkpoint directories and local caches were cleaned during experimentation.