Benchmark Recovery Does Not Certify a Frozen Tool-Trigger Contract After Quantization

Rui Li **Shuang Cao**
Hill Research

Abstract

Quantizing an action-bearing language model changes the scores that trigger tool calls, even when benchmark accuracy recovers. We audit this interface by freezing full-precision action labels, matching candidate training and threshold-retuning privileges, and evaluating held-out decision disagreement alongside externally labeled action errors. On Llama-3.1-70B and Qwen-2.5-72B at W4A16, standard-loss quantization-aware training plus threshold search recovers MMLU to within 0.3 points of FP16, while false tool-trigger rates at matched 90% recall remain 4.2 and 3.8 percentage points higher. A second controller with different actions and logistic-head scores reproduces this quality/action dissociation. Frontier-Aware Quantization (FAQ) combines boundary-sensitive precision allocation, a frontier-preservation loss, and threshold retuning; at the same 500-step budget it reduces frontier disagreement by 3.8/3.9 points and tool error by 2.1/2.0 points relative to the matched control. Separately labeled retrieval and stopping tests show corresponding action-level gains. Within FAQ, removing the frontier loss leaves Llama MMLU unchanged while raising disagreement by 3.6 points, isolating an improvement that benchmark quality misses. Repeated buffers establish the result's stability, while the precision sweep identifies substantially smaller action-error excesses at W8A8. These findings establish a concrete acceptance protocol that pairs recovered language-model quality with reference preservation and task-grounded action tests.

1 Introduction

A tool call is a discrete commitment: it can dispatch an external request, incur execution cost, or change the next step of an agent's workflow. Tool-using language models therefore expose an acceptance question that average benchmark accuracy does not resolve [Schick et al., 2023, Patil et al., 2024, Yao et al., 2023]. When compression changes a controller's scores, a small perturbation near a threshold can change an action even if the answer distribution remains accurate on a general language benchmark. The object to preserve is the deployed score-to-action interface.

We study this object with a freeze-and-retune audit. The full-precision controller's action labels remain fixed; each compressed candidate receives explicitly specified adaptation privileges, including held-out threshold search. A report-disjoint comparison then measures frontier crossing rate (FCR), the fraction of action decisions that disagree with the reference. Externally labeled action tests determine whether that disagreement also changes task behavior. The protocol fixes the object being preserved while varying the candidate's repair privileges, so each comparison answers a concrete acceptance question: whether quality recovery, probability calibration, or action-directed adaptation restores the interface.

The decisive comparison grants the compressed model both training and threshold adaptation. On Llama-3.1-70B-Instruct and Qwen-2.5-72B-Instruct at W4A16, standard-loss QAT plus threshold search reaches MMLU 78.9/82.0 against FP16 79.2/82.3. Yet BFCL false tool-trigger rates at

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

matched 90% recall increase from 8.2/7.6% to 12.4/11.4%. Thus a 0.3-point quality gap coexists with a 4.2/3.8-point action-error excess. The same pattern appears under a second controller with different actions, probability scores, and an independent audit buffer. W8A8 yields substantially smaller excesses, making precision an actionable part of the acceptance decision.

The contribution has three connected parts.

- A paired frozen-interface protocol that separates immutable reference labels, candidate calibration privileges, report-set preservation, and externally grounded action errors (Sections 3–4).
- A matched-control counterexample to benchmark-only acceptance: near-FP16 MMLU coexists with excess tool error across two controller designs. Repeated buffers, calibration comparisons, and precision sweeps characterize the separation, while labeled retrieval and stopping tests establish consequences beyond tool triggering (Section 5).
- A measured frontier-aware repair response: FAQ reduces FCR to 12.8/12.1% and tool FTR to 10.3/9.4% at the matched update budget, with $1.76\times$/$1.73\times$ FP16 throughput on the evaluated A100 stack. Component ablations identify the contribution of the frontier loss within this repair recipe.

2 Related work and positioning

GPTQ [Frantar et al., 2023], AWQ [Lin et al., 2024], SmoothQuant [Xiao et al., 2023], LLM.int8() [Dettmers et al., 2022], ZeroQuant [Yao et al., 2022], OmniQuant [Shao et al., 2024], QuIP# [Tseng et al., 2024], and AQLM [Egiazarian et al., 2024] develop complementary approaches to low-bit approximation and efficient inference. QAT and adaptation methods further recover task performance through training [Jacob et al., 2018, Nagel et al., 2022, Liu et al., 2024, Li et al., 2024, Dettmers et al., 2023]. GPTQ's weight reconstruction, AWQ's activation-aware weight protection, and standard-loss QAT address compression through different optimization targets. Here their outputs enter the same frozen action contract, with candidate threshold adaptation explicitly controlled. The retuned QAT/GPTQ comparison and FAQ loss ablation then test whether improving benchmark quality ranks interface preservation correctly (Table 3).

Calibration measures the agreement between probabilities and outcomes [Guo et al., 2017, Platt, 2000, Mukhoti et al., 2020, Kumar et al., 2019, Minderer et al., 2021]. Selective prediction and conformal methods connect uncertainty to coverage and decision rules [Geifman and El-Yaniv, 2017, Angelopoulos and Bates, 2023]. The distinctive constraint here is an immutable reference labeling coupled with candidate-only retuning. Threshold search can change FCR while leaving ECE and Brier scores exactly unchanged; Section 5.5 quantifies this separation. Probability calibration and reference preservation consequently supply complementary evidence for an action interface.

Toolformer, Gorilla, ReAct, and Reflexion place external actions inside language-model workflows [Schick et al., 2023, Patil et al., 2024, Yao et al., 2023, Shinn et al., 2023]. BFCL evaluates tool-use behavior [Patil et al., 2025]. We project its labels onto the tool-trigger frontier and pair that task test with an audit over the controller's full action set. Pairing BFCL with immutable controller labels connects two distinct evaluation objects: whether a compressed model chooses the reference action, and whether its action agrees with a task label. The cross-controller experiment tests this relationship under different action semantics, while the precision sweep makes its consequence visible for a compression choice.

3 Audit target: reference preservation and action error

Controller A exposes four direct logit-space margins from its controller wrapper: tool-call versus continue, retrieve versus continue, abstain versus answer, and stop/EOS versus continue. The same score definitions apply to every candidate and to the corresponding task projections. Let $s_k(x;\theta)$ denote the FP16 score and τ_k^{FP} its reference threshold. A quantized candidate uses parameters $\hat{\theta}$ and threshold $\hat{\tau}_k$. Define

$$d_k^{\mathrm{FP}}(x) = s_k(x;\theta) - \tau_k^{\mathrm{FP}}, \qquad \hat{d}_k(x) = s_k(x;\hat{\theta}) - \hat{\tau}_k. \tag{1}$$

The action triggers exactly when its distance is strictly positive; zero belongs to the non-trigger alternative. With $h(d) = \mathbb{1}[d > 0]$, report-set FCR is

$$\mathrm{FCR} = \frac{1}{K|\mathcal{D}^{\mathrm{rep}}|} \sum_{x \in \mathcal{D}^{\mathrm{rep}}} \sum_{k=1}^{K} \mathbb{1}\left[h(d_k^{\mathrm{FP}}(x)) \neq h(\hat{d}_k(x))\right]. \tag{2}$$

Controller A has $K = 4$ and 512 report turns, hence 2,048 paired decisions. Controller B has $K = 3$ and 768 report turns, hence 2,304 decisions; its FCR is reported separately. All displayed FCR values are percentages. Averaging frontiers is a preservation-accounting convention, with task tests providing action-specific consequences.

For a binary task label $y = 1$ indicating that an action is required, false-trigger rate is $\mathrm{FTR} = \mathrm{FP}/(\mathrm{FP} + \mathrm{TN})$ and recall is $\mathrm{TP}/(\mathrm{TP} + \mathrm{FN})$. BFCL tool FTR at 90% recall is a matched-recall operating characteristic on externally labeled turns. The task operating threshold is denoted $t_{\mathrm{task}}(0.90)$; the calibration-selected preservation threshold is $\hat{\tau}_k$. The former indexes the labeled score–recall curve, whereas the latter maximizes agreement with frozen reference decisions. Thus the two tests have different targets and populations. A threshold-only change leaves the matched-recall curve invariant for fixed scores and labels; its preservation effect is measured through FCR. A 4.2-point FTR excess corresponds to 42 additional triggers per 1,000 eligible no-tool turns at that recall. Abstention accuracy (AbsAcc) is agreement with the abstain/answer labels on TriviaQA-uncertain, a task-specific evaluation derived from TriviaQA [Joshi et al., 2017]; it uses its own evaluated-turn denominator. Retrieval-necessity and reference-boundary projections are reported separately in Appendix A.4.

FCR measures preservation of the reference decisions; external labels measure task correctness. These are complementary requirements because preserving an erroneous reference can preserve its error. A strictly increasing score transform with a correspondingly translated threshold leaves binary decisions and FCR invariant. It need not preserve numerical distances, a fixed-width boundary band, or the product loss used below.

4 Frozen-reference audit and frontier-aware repair

4.1 Split discipline and threshold adaptation

Controller A uses 2,048 ShareGPT-style conversation turns, partitioned into 1,024 training, 512 calibration, and 512 report turns. The same partition and reference labels are shared across methods. The standard 128-example PTQ layer-calibration set is separate from these action-audit roles. FAQ and standard-loss QAT update on audit-train; explicit threshold-search arms use audit-cal; report turns enter neither adaptation path. Within each comparison, the reference is frozen before candidate adaptation; the fresh-buffer experiment repeats this construction independently before freezing each new reference. This separates sensitivity to contract construction from candidate repair under a fixed contract (Appendix A.2).

The reference thresholds are fixed controller parameters, established before candidate comparison. In tool/retrieve/abstain/stop order, the large-model raw thresholds are $(0.62, 0.41, -0.18, 1.07)$ for Llama and $(0.55, 0.37, -0.24, 0.98)$ for Qwen. Candidate retuning maximizes agreement with these immutable FP16 labels, independently for each frontier. For each candidate frontier, calibration margins are standardized using calibration statistics; a 601-value grid from -3 to 3 in increments of 0.01 is mapped back to raw-score thresholds. Ties choose the larger threshold. Reference labels remain fixed throughout this optimization.

4.2 Frontier-sensitive allocation and training

FAQ combines layer precision allocation, a boundary-sensitive training objective, and final threshold retuning. The boundary-critical training subset is $\mathcal{D}_{\mathrm{bc}} = \{x \in \mathcal{D}^{\mathrm{train}} : \min_k |d_k^{\mathrm{FP}}(x)| < \epsilon\}$; all other training turns are interior. Controller A uses $\epsilon = 0.5$ in logit space. This partition affects FAQ training and localization, while Eq. 2 evaluates the full report set and contains no ϵ.

For each transformer block ℓ and candidate weight precision $b \in \{3, 4, 8\}$, a train-only sensitivity scan injects Gaussian block-output noise with variance matched to quantization error and estimates the resulting frontier-crossing cost $c_\ell(b)$. Dynamic programming over blocks and cumulative weight-bit

budget solves

$$\min_{b_1,\ldots,b_L} \sum_{\ell=1}^{L} c_\ell(b_\ell) \quad \text{subject to} \quad \sum_{\ell=1}^{L} |\mathbf{W}_\ell| b_\ell \le B, \quad b_\ell \in \{3,4,8\}. \tag{3}$$

The costs guide a global constrained allocation across transformer blocks under the shared weight-bit budget. W4A16, W4A8, and W8A8 specify the target average weight budget and activation precision. Uniform-weight baselines and FAQ therefore share a budget, while FAQ can distribute weight precision nonuniformly.

Starting from PTQ-initialized quantized weights, FAQ runs 500 AdamW updates with straight-through gradient estimates. It combines the standard language-model loss with

$$\mathcal{L} = \mathcal{L}_{\text{task}} + \lambda \mathcal{L}_{\text{frontier}}, \tag{4}$$

$$\mathcal{L}_{\text{frontier}} = \frac{1}{|\mathcal{D}_{\text{bc}}|} \sum_{x \in \mathcal{D}_{\text{bc}}} \sum_{k=1}^{K} \max\Big(0, m - d_k^{\text{FP}}(x)\tilde{d}_k(x)\Big), \tag{5}$$

where $\tilde{d}_k(x) = s_k(x;\hat{\theta}) - \tau_k^{\text{FP}}$ during training, $\lambda = 0.1$, and $m = 0.3$. The loss averages over selected examples and sums their frontiers. Consequently, selecting a boundary-critical turn does not mask its interior frontiers: a frontier contributes whenever its product is below m. After training, calibration-only threshold search supplies $\hat{\tau}_k$ for evaluation. This ordering separates differentiable score repair from held-out decision retuning.

Algorithm 1 Frontier-Aware Quantization (FAQ)

Require: FP16 model, fixed reference thresholds, disjoint train/cal/report sets, bit budget B
1: Freeze FP16 decisions; identify boundary-critical training turns
2: Estimate train-only layer costs $c_\ell(b)$ for $b \in \{3,4,8\}$
3: Solve Eq. 3; initialize quantized weights by PTQ
4: **for** $t = 1, \ldots, 500$ **do**
5: Sample training turns; update weights using Eqs. 4–5, AdamW and STE
6: **end for**
7: Select each candidate threshold on calibration data against fixed FP16 decisions
8: Evaluate report FCR and separate externally labeled task metrics
9: **return** Quantized model and candidate thresholds

5 Experiments

5.1 Models, comparators, and measurement

We evaluate Llama-3.1-8B/70B-Instruct [Grattafiori et al., 2024] and Qwen-2.5-7B/72B-Instruct [Qwen Team, 2024]. All weight quantization uses group size 128. The main W4A16 comparison includes GPTQ, AWQ, QuIP#, standard-loss QAT, GPTQ plus threshold search, QAT plus threshold search, and FAQ. Unretuned and retuned pairs measure threshold adaptation through FCR. For these pairs, the matched-recall task metric is reported once, in the retuned row. Standard-loss QAT uses uniform W4 weights without FAQ's boundary stratification or mixed precision, for the same 500 AdamW+STE steps. The crossed QAT+τ control adds the identical calibration-only sweep. It therefore matches both training budget and retuning privilege.

MMLU uses five-shot evaluation [Hendrycks et al., 2021]; tool FTR uses BFCL labels [Patil et al., 2025]. FCR, FTR and AbsAcc in Table 1 retain the three-run means and standard deviations for seeds 13/21/34. Report turns and reference labels are shared between candidate arms. Separate buffer-level ranges and repartition SDs are identified explicitly in Appendix A.2; they are different sources of variation from run SD. MMLU and throughput are point estimates. All timing uses 8×NVIDIA A100-80GB, CUDA 12.4, PyTorch 2.3, batch size 32, and sequence length 2,048. Threshold search changes no language-model weights, so its parent model's MMLU and token-generation timing apply to the threshold-only arm.

Table 1: W4A16 quality, action preservation, and task errors. Rates are percentages; ± is three-run SD. $+\tau$ denotes threshold search. MMLU and throughput are point estimates; dashes avoid repeating matched-recall FTR within threshold-only pairs.

Method	FCR ↓	FTR ↓	AbsAcc ↑	MMLU	Tput
Llama-3.1-70B					
FP16 (reference)	—	8.2±0.3	81.4±0.5	79.2	1.00×
GPTQ	23.7±1.1	—	73.1±0.9	78.4	1.82×
AWQ	21.4±0.9	15.1±0.7	74.3±0.8	78.6	1.84×
QuIP#	19.8±1.0	14.3±0.6	75.0±0.7	78.7	1.71×
QAT	17.3±0.8	—	76.2±0.6	78.9	1.83×
GPTQ + τ	16.1±0.7	11.9±0.5	76.8±0.5	78.4	1.82×
QAT + τ	16.6±0.7	12.4±0.5	76.4±0.5	78.9	1.83×
FAQ	12.8±0.6	10.3±0.4	78.9±0.4	79.0	1.76×
Qwen-2.5-72B					
FP16 (reference)	—	7.6±0.3	82.8±0.4	82.3	1.00×
GPTQ	22.1±1.0	—	75.4±0.8	81.5	1.79×
AWQ	20.3±0.9	13.8±0.6	76.2±0.7	81.7	1.81×
QuIP#	18.9±0.8	13.1±0.5	76.9±0.6	81.8	1.68×
QAT	16.7±0.7	—	78.1±0.5	82.0	1.80×
GPTQ + τ	15.4±0.6	10.9±0.4	78.6±0.5	81.5	1.79×
QAT + τ	16.0±0.6	11.4±0.4	78.4±0.5	82.0	1.80×
FAQ	12.1±0.5	9.4±0.3	80.7±0.4	82.1	1.73×

5.2 Quality recovery and interface recovery separate

QAT+τ strengthens the quality/action dissociation (Table 1). Its MMLU deficit is 0.3 points on both large models, while FTR exceeds FP16 by 4.2/3.8 points, with paired 95% intervals [3.4, 5.0]/[3.1, 4.5] (Table 2). GPTQ+τ has a 0.8-point MMLU deficit and 3.7/3.3-point FTR excess. Figure 1 places these two gaps alongside FAQ. The aggregate four-frontier disagreement is 16.6/16.0% for QAT+τ and 12.8/12.1% for FAQ. The smaller checkpoints exhibit the same direction with greater compression sensitivity (Appendix A.7).

FAQ improves FCR by 3.8/3.9 points and FTR by 2.1/2.0 points against the crossed control. The paired 95% intervals for FAQ-minus-control FCR are $[-4.5, -3.1]$/$[-4.6, -3.2]$ points. Against GPTQ+τ, its FCR improvement is 3.3 points on both models. FAQ's residual tool-FTR excess is 2.1/1.8 points: the audit measures both the action errors removed by repair and the remaining distance to FP16.

Table 2: Matched-control contrasts at W4A16. Differences and 95% intervals are in percentage points; intervals are distinct from the run SDs in Table 1.

Contrast	Llama-3.1-70B	Qwen-2.5-72B
QAT+τ minus FP16 tool FTR	4.2 [3.4, 5.0]	3.8 [3.1, 4.5]
FAQ minus QAT+τ FCR	-3.8 $[-4.5, -3.1]$	-3.9 $[-4.6, -3.2]$
FAQ minus QAT+τ tool FTR	-2.1	-2.0

Table 3 explains how different repairs affect the acceptance criteria. Adding the frontier loss within FAQ changes Llama FCR from 16.4 to 12.8% at identical MMLU 79.0. Threshold retuning changes FCR while preserving the model's probabilities and calibration scores. The between-method point estimates provide a further selection example: retuned QAT gains 0.5 MMLU points over retuned GPTQ on both large models, yet has 0.5/0.6-point higher FCR. Together, the loss ablation and threshold intervention show directly why recovered benchmark or calibration scores leave an independent preservation question.

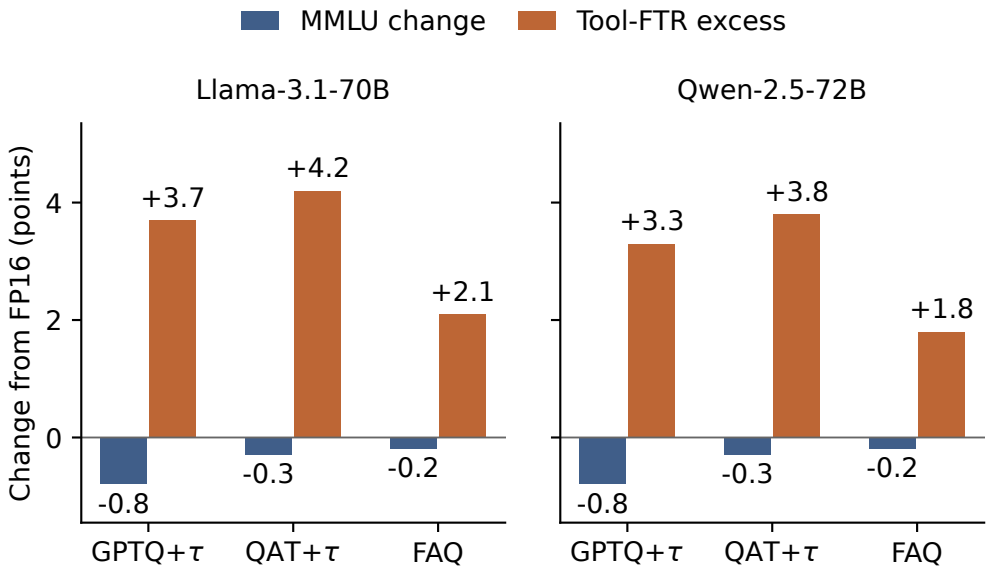

Figure 1: Benchmark quality and tool-action acceptance at W4A16. Bars show MMLU change and BFCL tool-FTR excess relative to each model's FP16 reference; labels are percentage-point differences computed from Table 1. MMLU has no uncertainty estimate. Near-zero quality loss coexists with substantial excess tool firing, including after matched QAT and threshold retuning.

Table 3: Interventions that separate acceptance criteria. Entries reuse Tables 1, 7, and 11; arrows connect the compared arms. FCR entries are run means; MMLU and ECE/Brier entries are point estimates or displayed means as specified in the source tables. Rates are percentages, ECE/Brier use a 0–1 scale. Llama is shown unless both models are named.

Controlled comparison	Quality or calibration response	Reference-preservation response
GPTQ+τ to QAT+τ; same retuning privilege	MMLU 78.4→78.9 (Llama), 81.5→82.0 (Qwen)	FCR 16.1→16.6 and 15.4→16.0
FAQ without frontier loss to full FAQ; other components fixed	MMLU 79.0→79.0	FCR 16.4→12.8
GPTQ to GPTQ+τ; model scores fixed	ECE 0.058→0.058; Brier 0.131→0.131	FCR 23.7→16.1

5.3 Replication across controller semantics

Controller B uses `api_call`, `clarify`, and `escalate` against `answer_directly`. Scores come from a frozen temperature-calibrated logistic head; FP16 thresholds target action precision 0.80/0.70/0.90. Its independent 3,072-turn buffer, with no source conversations shared with Controller A, is split 1,536/768/768 into train/cal/report. Candidate thresholds are retuned against frozen FP16 decisions on calibration data only.

Table 4 reproduces the action-error increase with these different scores and actions. QAT+τ raises `api_call` FTR from 6.9 to 10.7% on Llama and from 6.2 to 9.8% on Qwen. FAQ reduces these rates to 8.9/8.2% and three-frontier FCR to 11.5/10.8%. The changed action vocabulary, logistic score map, precision-target reference, and independent conversations jointly test whether the quality/action separation transfers beyond Controller A. Each controller retains its own FCR denominator.

Table 4: Controller B at W4A16. Percentages are reported point estimates. FCR_B averages three frontiers over 768 report turns; FTR evaluates the externally labeled `api_call` action.

	Llama-3.1-70B		Qwen-2.5-72B	
Method	FCR_B	FTR	FCR_B	FTR
FP16	0	6.9	0	6.2
GPTQ+τ	14.7	10.4	13.9	9.6
QAT+τ	15.2	10.7	14.3	9.8
FAQ	11.5	8.9	10.8	8.2

5.4 Retrieval and stopping consequences

The action tests extend beyond tool triggering. Table 5 evaluates retrieval false triggers using retrieval-necessity labels and premature stopping using reference turn boundaries. Relative to QAT+τ, FAQ reduces retrieval false triggers by 2.6/2.3 percentage points and premature stopping by 1.6/1.5 points on Llama/Qwen. The FP16 row has the lowest error on both projections, and FAQ improves both over each retuned control. These are direct task-specific comparisons alongside the aggregate preservation audit.

Retrieval eligibility covers 61.3% of its task population and stopping eligibility 78.4%; each metric conditions on its own labeled subset (Appendix A.4). This distinction matters operationally: a changed reference threshold can retain FAQ's 3.3-point FCR advantage over GPTQ+τ while worsening the reference policy's tool decisions. Pairing task labels with preservation therefore evaluates both the quality of the reference and the fidelity of the compressed interface.

Table 5: Controller A task-grounded projections at W4A16 and matched 90% recall. Entries are point estimates in percent. Retrieval false-trigger rate and premature-stop rate retain distinct eligible populations.

	Llama-3.1-70B		Qwen-2.5-72B	
Method	Retrieval	Stop	Retrieval	Stop
FP16	9.6	5.2	8.9	4.7
GPTQ+τ	13.4	7.8	12.3	7.1
QAT+τ	13.9	8.1	12.8	7.4
FAQ	11.3	6.5	10.5	5.9

5.5 Probability calibration and preserved decisions

Threshold-free calibration metrics cannot register threshold-only repairs. On Llama, GPTQ and GPTQ+τ share ECE 0.058 ± 0.004 and Brier 0.131 ± 0.005, while FCR falls from 23.7 to 16.1%. QAT and QAT+τ likewise share ECE 0.048 and Brier 0.124 while FCR changes from 17.3 to 16.6%. These comparisons use the same tool probabilities and BFCL labels within each pair. The ECE-optimized temperature arm further achieves ECE/Brier 0.037/0.117 versus FP16 0.041/0.118, while its fixed-contract FCR is $17.9 \pm 0.8\%$; Qwen yields ECE/Brier 0.034/0.109 and FCR 17.0% (Appendix A.3). Better probability calibration therefore coexists with reference-decision disagreement. Positive scalar temperature preserves ranking, so this observation concerns a fixed threshold contract, not a change in the matched-recall operating curve.

5.6 Buffer and boundary sensitivity

Five fresh disjoint 2,048-turn buffers preserve the ordering of GPTQ+τ, QAT+τ, and FAQ in every buffer. Their mean FCRs are 16.3, 16.8, and 13.0%, with respective ranges 15.4–17.4, 15.9–17.9, and 12.1–14.0%. Ten calibration/report repartitions give a FAQ-minus-GPTQ+τ gap of -3.3 points (SD 0.3), with the weakest gap -2.6. Increasing total buffer size from 2,048 to 8,192 changes FCR by 0.2 points for both methods. Retraining FAQ at $\epsilon \in \{0.10, 0.25, 0.50, 1.00\}$ preserves its advantage on both large models, with a nonmonotone optimum near the default band (Appendix A.2).

5.7 Precision and prompt shifts

Precision distinguishes persistent from small action error (Table 6). Retuned GPTQ's W4A8 tool-FTR excess is 5.5/5.2 points, alongside a larger MMLU decline than W4A16. At W8A8, the excess shrinks to 0.5/0.3 points and the reported intervals span zero; FAQ does not separate from the retuned W8A8 controls. This precision sweep distinguishes the near-quality W4A16 counterexample, the larger quality/action losses of W4A8, and W8A8's small task-error excess. W8A8 thus supplies a favorable operating comparison; its uncertainty interval quantifies the residual error for comparison with an application's tolerance.

The four prompt settings in Appendix A.5 retain the reference score definitions and thresholds while changing the evaluation inputs. GPTQ+τ FCR rises from 16.1% in distribution to 20.3/21.6/19.1/22.4% on WildChat, tool-schema, task-dialogue, and non-English shifts. FAQ rises from 12.8% to 15.6/16.9/14.9/17.8%; it retains lower disagreement in each setting.

Table 6: Tool-FTR excess over FP16 for GPTQ+τ, in percentage points. W8A8 brackets are reported paired 95% intervals. The W8A8 intervals quantify the residual task-error excess; FCR remains a separate preservation measure.

Precision	Llama-3.1-70B	Qwen-2.5-72B
W4A16	3.7	3.3
W4A8	5.5	5.2
W8A8	0.5 [−0.1, 1.1]	0.3 [−0.3, 0.9]

5.8 Repair components and inference cost

Table 8 aligns adaptation access with the matched budget. On Llama, FAQ delivers 4,078 tokens/s against GPTQ+τ's 4,218 tokens/s, a 3.3% reduction, while lowering FCR from 16.1 to 12.8%. Table 7 separates interventions within FAQ: removing the frontier loss raises FCR to 16.4%; removing mixed precision to 14.9%; removing stratification to 14.1%; and omitting final retuning to 13.6%. The loss ablation keeps FAQ's other components, whereas the crossed control uses uniform W4 and ordinary sampling. Their respective 16.4 and 16.6% FCR values reflect distinct interventions. The 3.6-point loss-ablation effect at unchanged MMLU identifies the contribution of action-directed optimization within FAQ.

Table 7: FAQ ablation on Llama-3.1-70B W4A16. Each row removes the named component; the final row is plain QAT. ΔFCR is the percentage-point increase over FAQ. Dashes omit repeated matched-recall task rates.

Variant	FCR ↓	Δ FCR	FTR ↓	MMLU
Full FAQ	12.8±0.6	—	10.3±0.4	79.0
− Frontier loss ($\lambda = 0$)	16.4±0.8	+3.6	12.5±0.5	79.0
− Mixed precision (uniform W4)	14.9±0.7	+2.1	11.4±0.4	78.7
− Audit buffer stratification	14.1±0.7	+1.3	11.0±0.4	78.9
− Post-quant τ-calibration	13.6±0.6	+0.8	—	79.0
− All components (plain QAT)	17.3±0.8	+4.5	—	78.9

6 Discussion

A compression choice combines benchmark loss, per-action error excess, reference disagreement, and throughput, each with its own application tolerance. The loss and threshold interventions explain why these coordinates must be measured separately: frontier loss improves FCR at unchanged MMLU, and threshold search changes FCR at unchanged ECE/Brier. Each reveals a decision improvement that its unchanged quality or calibration metric would miss. At W4A16, FAQ trades 3.3% of retuned GPTQ's Llama throughput for a 3.3-point FCR reduction; W8A8 supplies a lower-error precision

alternative. Separately labeled retrieval and stop tests extend this decision beyond the tool frontier (Section 5.4).

The evaluated scope comprises two scalar-threshold controller designs and four instruction-tuned checkpoints, with the extended controls on the two large models. Stress-buffer rates describe paired preservation under the specified sampling protocol; fresh-buffer experiments characterize their stability. Retrieval and stopping tests use eligible task-proxy subsets, and throughput describes the specified A100 configuration. Improved preservation can reduce unintended action changes while also preserving an erroneous reference policy, which is why task-grounded acceptance remains essential.

7 Conclusion

Benchmark recovery and action-interface recovery are empirically distinct outcomes of language-model quantization. With standard-loss QAT and held-out threshold retuning, the large W4A16 models recover MMLU to within 0.3 points of FP16 while retaining 4.2/3.8-point tool-error excesses at matched recall. A second controller reproduces the pattern, and repeated buffers establish its stability. FAQ improves reference preservation and separately labeled tool, retrieval, and stopping behavior under matched adaptation budgets; its loss ablation demonstrates an action-directed gain at unchanged MMLU. The precision comparison identifies W8A8 as a lower-error operating regime. These results turn the frozen-reference audit into a practical candidate-selection protocol: evaluate language-model quality, preserved decisions, task-grounded action errors, and inference cost together.

Acknowledgments and Disclosure of Funding

This work received no external funding. The authors declare no competing interests.

References

Anastasios N Angelopoulos and Stephen Bates. Conformal prediction: A gentle introduction. *Foundations and Trends in Machine Learning*, 16(4):494–591, 2023. doi: 10.1561/2200000101. URL https://doi.org/10.1561/2200000101.

Tim Dettmers, Mike Lewis, Younes Belkada, and Luke Zettlemoyer. LLM.int8(): 8-bit matrix multiplication for transformers at scale. In *Advances in Neural Information Processing Systems*, volume 35, pages 30318–30332, 2022. doi: 10.52202/068431-2198. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/c3ba4962c05c49636d4c6206a97e9c8a-Paper-Conference.pdf.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. QLoRA: Efficient finetuning of quantized LLMs. In *Advances in Neural Information Processing Systems*, volume 36, pages 10088–10115, 2023. doi: 10.52202/075280-0441. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/1feb87871436031bdc0f2beaa62a049b-Abstract.html.

Vage Egiazarian, Andrei Panferov, Denis Kuznedelev, Elias Frantar, Artem Babenko, and Dan Alistarh. Extreme compression of large language models via additive quantization. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235, pages 12284–12303, 2024. URL https://proceedings.mlr.press/v235/egiazarian24a.html.

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

Yonatan Geifman and Ran El-Yaniv. Selective classification for deep neural networks. In *Advances in Neural Information Processing Systems*, volume 30, 2017. URL https://proceedings.neurips.cc/paper/2017/hash/4a8423d5e91fda00bb7e46540e2b0cf1-Abstract.html.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, et al. The Llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024. doi: 10.48550/arXiv.2407.21783. URL https://arxiv.org/abs/2407.21783.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. On calibration of modern neural networks. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70, pages 1321–1330, 2017. URL https://proceedings.mlr.press/v70/guo17a.html.

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

Benoit Jacob, Skirmantas Kligys, Bo Chen, Menglong Zhu, Matthew Tang, Andrew Howard, Hartwig Adam, and Dmitry Kalenichenko. Quantization and training of neural networks for efficient integer-arithmetic-only inference. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2704–2713, 2018. URL https://openaccess.thecvf.com/content_cvpr_2018/html/Jacob_Quantization_and_Training_CVPR_2018_paper.html.

Mandar Joshi, Eunsol Choi, Daniel S. Weld, and Luke Zettlemoyer. TriviaQA: A large scale distantly supervised challenge dataset for reading comprehension. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1601–1611, 2017. doi: 10.18653/v1/P17-1147. URL https://aclanthology.org/P17-1147/.

Ananya Kumar, Percy S. Liang, and Tengyu Ma. Verified uncertainty calibration. In *Advances in Neural Information Processing Systems*, volume 32, 2019. URL https://proceedings.neurips.cc/paper_files/paper/2019/hash/f8c0c968632845cd133308b1a494967f-Abstract.html.

Yixiao Li, Yifan Yu, Chen Liang, Pengcheng He, Nikos Karampatziakis, Weizhu Chen, and Tuo Zhao. LoftQ: LoRA-fine-tuning-aware quantization for large language models. In *International Conference on Learning Representations*, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/file/39ec972afab01e0d8ddc6834a9d12ac1-Paper-Conference.pdf.

Ji Lin, Jiaming Tang, Haotian Tang, Shang Yang, Wei-Ming Chen, Wei-Chen Wang, Guangxuan Xiao, Xingyu Dang, Chuang Gan, and Song Han. AWQ: Activation-aware weight quantization for on-device LLM compression and acceleration. In *Proceedings of Machine Learning and Systems*, volume 6, pages 87–100, 2024. URL https://proceedings.mlsys.org/paper_files/paper/2024/hash/42a452cbafa9dd64e9ba4aa95cc1ef21-Abstract-Conference.html.

Zechun Liu, Barlas Oguz, Changsheng Zhao, Ernie Chang, Pierre Stock, Yashar Mehdad, Yangyang Shi, Raghuraman Krishnamoorthi, and Vikas Chandra. LLM-QAT: Data-free quantization aware training for large language models. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 467–484, 2024. doi: 10.18653/v1/2024.findings-acl.26. URL https://aclanthology.org/2024.findings-acl.26/.

Matthias Minderer, Josip Djolonga, Rob Romijnders, Frances Hubis, Xiaohua Zhai, Neil Houlsby, Dustin Tran, and Mario Lucic. Revisiting the calibration of modern neural networks. In *Advances in Neural Information Processing Systems*, volume 34, pages 15682–15694, 2021. URL https://proceedings.neurips.cc/paper/2021/hash/8420d359404024567b5aefda1231af24-Abstract.html.

Jishnu Mukhoti, Viveka Kulharia, Amartya Sanyal, Stuart Golodetz, Philip Torr, and Puneet Dokania. Calibrating deep neural networks using focal loss. In *Advances in Neural Information Processing Systems*, volume 33, pages 15288–15299, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/aeb7b30ef1d024a76f21a1d40e30c302-Abstract.html.

Markus Nagel, Marios Fournarakis, Yelysei Bondarenko, and Tijmen Blankevoort. Overcoming oscillations in quantization-aware training. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162, pages 16318–16330, 2022. URL https://proceedings.mlr.press/v162/nagel22a.html.

Shishir G. Patil, Tianjun Zhang, Xin Wang, and Joseph E. Gonzalez. Gorilla: Large language model connected with massive APIs. In *Advances in Neural Information Processing Systems*, volume 37, pages 126544–126565, 2024. doi: 10.52202/079017-4020. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/e4c61f578ff07830f5c37378dd3ecb0d-Abstract-Conference.html.

Shishir G Patil, Huanzhi Mao, Fanjia Yan, Charlie Cheng-Jie Ji, Vishnu Suresh, Ion Stoica, and Joseph E. Gonzalez. The berkeley function calling leaderboard (BFCL): From tool use to agentic evaluation of large language models. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267, pages 48371–48392, 2025. URL https://proceedings.mlr.press/v267/patil25a.html.

John C. Platt. Probabilities for SV machines. In Alexander J. Smola, Peter Bartlett, Bernhard Schölkopf, and Dale Schuurmans, editors, *Advances in Large-Margin Classifiers*, pages 61–74. MIT Press, 2000. doi: 10.7551/mitpress/1113.003.0008. URL https://doi.org/10.7551/mitpress/1113.003.0008.

Qwen Team. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*, 2024. doi: 10.48550/arXiv.2412.15115. URL https://arxiv.org/abs/2412.15115.

Timo Schick, Jane Dwivedi-Yu, Roberto Dessì, Roberta Raileanu, Maria Lomeli, Eric Hambro, Luke Zettlemoyer, Nicola Cancedda, and Thomas Scialom. Toolformer: Language models can teach themselves to use tools. In *Advances in Neural Information Processing Systems*, volume 36, pages 68539–68551, 2023. doi: 10.52202/075280-2997. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/d842425e4bf79ba039352da0f658a906-Abstract-Conference.html.

Wenqi Shao, Mengzhao Chen, Zhaoyang Zhang, Peng Xu, Lirui Zhao, Zhiqian Li, Kaipeng Zhang, Peng Gao, Yu Qiao, and Ping Luo. OmniQuant: Omnidirectionally calibrated quantization for large language models. In *International Conference on Learning Representations*, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/c6483c8a68083af3383f91ee0dc6db95-Abstract-Conference.html.

Noah Shinn, Federico Cassano, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. Reflexion: Language agents with verbal reinforcement learning. In *Advances in Neural Information Processing Systems*, volume 36, pages 8634–8652, 2023. doi: 10.52202/075280-0377. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/1b44b878bb782e6954cd888628510e90-Abstract-Conference.html.

Albert Tseng, Jerry Chee, Qingyao Sun, Volodymyr Kuleshov, and Christopher De Sa. QuIP#: Even better LLM quantization with hadamard incoherence and lattice codebooks. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235, pages 48630–48656, 2024. URL https://proceedings.mlr.press/v235/tseng24a.html.

Guangxuan Xiao, Ji Lin, Mickael Seznec, Hao Wu, Julien Demouth, and Song Han. SmoothQuant: Accurate and efficient post-training quantization for large language models. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202, pages 38087–38099, 2023. URL https://proceedings.mlr.press/v202/xiao23c.html.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. ReAct: Synergizing reasoning and acting in language models. In *International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=WE_vluYUL-X.

Zhewei Yao, Reza Yazdani Aminabadi, Minjia Zhang, Xiaoxia Wu, Conglong Li, and Yuxiong He. Zeroquant: Efficient and affordable post-training quantization for large-scale transformers. In *Advances in Neural Information Processing Systems*, volume 35, pages 27168–27183, 2022. doi: 10.52202/068431-1970. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/adf7fa39d65e2983d724ff7da57f00ac-Abstract-Conference.html.

Wenting Zhao, Xiang Ren, Jack Hessel, Claire Cardie, Yejin Choi, and Yuntian Deng. WildChat: 1m ChatGPT interaction logs in the wild. In *International Conference on Learning Representations*, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/9421261e06f1a63a352b068f1ac90609-Abstract-Conference.html.

A.1 Matched adaptation and component effects

Table 8: Llama-3.1-70B W4A16 adaptation access and inference cost. Every arm excludes audit-report turns from adaptation. PTQ layer calibration is separate from audit calibration. QAT+τ reuses QAT's token-generation timing because only thresholds change.

Method	Steps	Retune	Weight allocation / loss	tok/s	FCR
GPTQ	0	No	Uniform W4 / PTQ	4,218	23.7
AWQ	0	No	Uniform W4 / PTQ	4,264	21.4
GPTQ+τ	0	Yes	Uniform W4 / PTQ	4,218	16.1
QAT	500	No	Uniform W4 / LM	4,241	17.3
QAT+τ	500	Yes	Uniform W4 / LM	4,241	16.6
FAQ, $\lambda = 0$	500	Yes	Mixed / LM, stratified	—	16.4
FAQ	500	Yes	Mixed / LM + frontier, stratified	4,078	12.8

Threshold search reduces Llama GPTQ FCR by 7.6 points, QAT by 0.7, and FAQ by 0.8; corresponding Qwen reductions are 6.7, 0.7, and 0.8. Training and threshold retuning therefore address different portions of the gap. The full-FAQ versus zero-frontier-loss comparison keeps stratification, allocation, and final retuning in place, while the standard-loss crossed control matches the update and threshold privileges with uniform precision and ordinary sampling. At MMLU 79.0 in both FAQ arms, the loss ablation changes FCR by 3.6 points and FTR by 2.2 points. This within-recipe comparison is the equal-quality evidence in Table 3; it complements the between-recipe crossed control.

A.2 Audit sensitivity

Fresh-buffer experiments rebuild disjoint 2,048-turn buffers, refit reference thresholds, and rerun adaptation and evaluation. Five such buffers yield GPTQ+τ mean FCR 16.3% (range 15.4–17.4), QAT+τ 16.8% (15.9–17.9), and FAQ 13.0% (12.1–14.0). The ordering holds in all five, with tool-FTR excesses of 3.0–4.3 points for the retuned GPTQ comparison. These ranges describe variation across buffers.

Ten calibration/report repartitions yield GPTQ+τ 16.2% (SD 0.6), FAQ 12.9% (SD 0.5), and a paired difference of -3.3 points (SD 0.3). Replacing the calibration pool with disjoint WildChat examples gives 17.4% versus 13.6%. This calibration-pool experiment changes adaptation input, whereas the prompt-shift test below fixes thresholds and changes only evaluation input.

Table 9: Llama-3.1-70B W4A16 buffer-size sensitivity. N counts all audit turns; FCR is evaluated on the report partition, not on all N turns. Values are percentages.

Total buffer N	512	1,024	2,048	4,096	8,192
GPTQ+τ FCR	16.8	16.4	16.1	16.0	15.9
FAQ FCR	13.5	13.1	12.8	12.7	12.6

Table 10: Boundary-band sensitivity at W4A16. FAQ is retrained for each ϵ; all cells are percentages. The retuned GPTQ comparison is unchanged because its training and the full-report metrics do not depend on ϵ.

Model	Metric	$\epsilon = 0.10$	0.25	0.50	1.00
Llama-70B	FAQ FCR	13.9	13.1	12.8	13.2
	FAQ FTR	11.0	10.6	10.3	10.5
Qwen-72B	FAQ FCR	13.1	12.4	12.1	12.5
	FAQ FTR	10.0	9.6	9.4	9.7

Across the band sweep, GPTQ+τ remains at FCR 16.1/15.4% and FTR 11.9/10.9%, and QAT+τ remains at FCR 16.6/16.0%. Llama's boundary-critical mass grows from 9.4% at $\epsilon = 0.10$ to 63.8%

at $\epsilon = 1.00$. The narrowest band supplies fewer boundary-critical training examples, and FAQ's response is nonmonotone as the band widens (Figure 2). These are Controller A logit-space bands; Controller B's probability scores define a different score space.

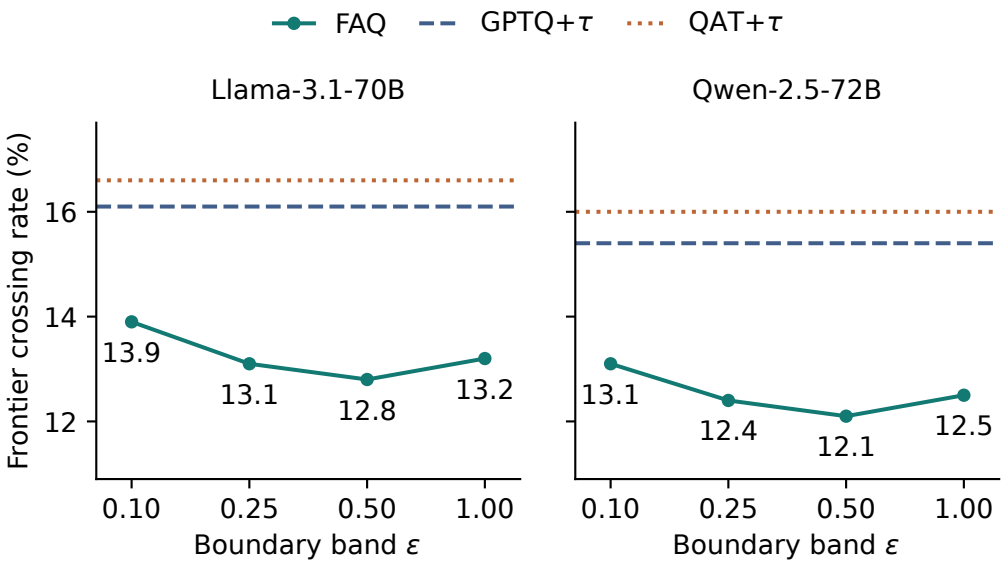

Figure 2: Boundary-band sensitivity. FAQ's FCR varies with its training band but remains below the unchanged GPTQ+τ and QAT+τ controls on both large models. Values correspond to Table 10; lines connect evaluated settings.

A.3 Calibration and frozen-contract decisions

Brier score averages squared probability error against binary tool labels. ECE aggregates the discrepancy between predicted probability and observed tool frequency within probability bins. These metrics use BFCL-labeled tool turns, whereas the FCR column evaluates the frozen controller report set. Within GPTQ/GPTQ+τ and QAT/QAT+τ, probabilities and labels are identical; threshold-only retuning changes neither calibration score. The temperature arm optimizes ECE and evaluates the resulting scores against the fixed reference contract. Translating a threshold along with a positive monotone temperature transform would instead preserve decisions, and the score ranking alone determines matched-recall operating characteristics.

Table 11: Llama-3.1-70B calibration and action preservation. ECE and Brier are on a 0–1 scale; FCR is a percentage. Uncertainty is run SD where reported. Temperature calibration can improve probability metrics while leaving substantial fixed-contract disagreement.

Arm	ECE	Brier	FCR
FP16	0.041	0.118	0
GPTQ	0.058 ± 0.004	0.131 ± 0.005	23.7 ± 1.1
GPTQ+τ	0.058 ± 0.004	0.131 ± 0.005	16.1 ± 0.7
QAT	0.048	0.124	17.3 ± 0.8
QAT+τ	0.048	0.124	16.6 ± 0.7
Temperature calibration	0.037	0.117	17.9 ± 0.8

The corresponding Qwen temperature arm reaches ECE 0.034, Brier 0.109, and FCR 17.0%. The invariance of threshold-free scores within each threshold-only pair and the temperature arm's residual disagreement give direct, intervention-based evidence of the distinction between probability calibration and reference preservation.

A.4 Task-grounded retrieval and stopping

The retrieval projection uses a retrieval-necessity proxy: the positive class comprises turns requiring retrieval, and false triggers are evaluated on eligible negative turns at 90% positive recall. The stopping projection uses reference turn boundaries; a trigger before the labeled stopping boundary is premature. Eligibility coverage is 61.3% for retrieval and 78.4% for stopping in their task-projection populations. Each rate is conditional on its own eligible labeled population, separate from the 2,048 frontier decisions in Controller A's FCR denominator. The proxies evaluate these action-specific tasks alongside direct tool labels and abstention labels.

Table 5 reports the resulting task rates in the main text. Matching recall permits an action-specific comparison at the same positive-class sensitivity; the eligibility coverage identifies the population to which each comparison applies.

A deliberately mis-specified FP16 tool threshold provides a reference-policy test. FAQ retains a 3.3-point FCR advantage over GPTQ+τ under this changed reference, even though the reference induces worse tool decisions. This result isolates the preservation target: matching the reference more closely can coexist with inheriting its policy error. External task labels are therefore evaluated alongside FCR, rather than inferred from it.

A.5 Prompt shifts and precision costs

The prompt-shift tests freeze the score definitions and calibrated thresholds from the in-distribution controller. WildChat [Zhao et al., 2024] changes the prompt source; the additional settings change tool schemas, task-oriented dialogue, and prompt language. Table 12 compares retuned GPTQ and FAQ under each setting. The WildChat comparison additionally includes GPTQ and AWQ with run SDs in Table 13.

Table 12: Llama-3.1-70B W4A16 FCR under frozen-threshold prompt shifts (percent). Both methods use the same report prompts within each setting.

Evaluation setting	GPTQ+τ	FAQ
In distribution	16.1	12.8
WildChat	20.3	15.6
Tool-schema shift	21.6	16.9
Task-oriented dialogue	19.1	14.9
Non-English prompts	22.4	17.8

Table 13: Prompt-distribution shift on Llama-3.1-70B W4A16. The tool/retrieve/abstain/stop score definitions and thresholds are frozen from the ShareGPT-built audit protocol; only the held-out evaluation conversations are replaced with WildChat prompts [Zhao et al., 2024].

Method	FCR (in-dist) ↓	FCR (WildChat) ↓	Degradation
GPTQ	23.7±1.1	26.4±1.3	+2.7
AWQ	21.4±0.9	24.1±1.1	+2.7
GPTQ + τ-search	16.1±0.7	20.3±0.9	+4.2
FAQ	12.8±0.6	15.6±0.8	+2.8

Table 14 connects precision with throughput and preservation. W8A8 lowers FCR for both GPTQ and FAQ relative to W4A8, while W4A8 offers higher throughput. These point estimates complement the matched-recall task comparison in Table 6, which evaluates a different acceptance criterion.

A.6 Evidence map for the action interface

The protected object is the FP16 score-to-action labeling. The interventions are quantization, model adaptation, and candidate threshold retuning. The audit boundary excludes report turns from all adaptation and keeps reference semantics fixed. The intervention contrasts in Table 3 distinguish weight

Table 14: Additional precision/throughput point estimates. These GPTQ and FAQ measurements complement the retuned-control tool-error comparison in Table 6; they are not seed-averaged comparisons against retuned controls.

Model	Method-bitwidth	Tput / FP16	FCR ↓
Llama-3.1-70B	GPTQ W4A8	1.95×	27.1
Llama-3.1-70B	GPTQ W8A8	1.42×	10.4
Llama-3.1-70B	FAQ W4A8	1.87×	14.7
Llama-3.1-70B	FAQ W8A8	1.39×	8.7
Qwen-2.5-72B	GPTQ W4A8	1.92×	25.8
Qwen-2.5-72B	GPTQ W8A8	1.39×	9.8
Qwen-2.5-72B	FAQ W4A8	1.84×	13.9
Qwen-2.5-72B	FAQ W8A8	1.36×	7.1

repair, threshold repair, and probability calibration by the measurements each changes. Table 15 connects each measured property to its comparator, held-out setting, and observed consequence.

Table 15: Interface evidence map. All rows assume fixed reference labels and disjoint adaptation/report roles. Preservation and externally labeled action error remain separate properties.

Object / perturbation	Enforcement and comparator	Held-out property / utility	Observed failure and evidence
Controller A / W4	Freeze four margins; QAT+τ matches updates and retuning	Four-frontier FCR; BFCL FTR; MMLU; A100 throughput	Recovered quality with action error (Tables 1, 8)
Controller B / W4	Freeze logistic head and precision-target labels; QAT+τ	Independent three-frontier report set; labeled API calls	Pattern transfers across action semantics (Table 4)
Tool probability / calibration	Freeze reference decisions; compare threshold-only and temperature arms	ECE/Brier and FCR	Probability quality and action preservation separate (Table 11)
Precision / prompt changes	Retuned GPTQ; same reference within each setting	Tool error by precision; prompt-shift FCR	Small W8 excess; larger shift disagreement (Tables 6, 12)
Reference policy	Change FP16 tool threshold; compare FAQ with GPTQ+τ	Preservation under the changed reference	Lower FCR can preserve policy error (Appendix A.4)
Retrieve / stop	Separate task-proxy labels; QAT+τ	Eligible-population action errors	Partial task recovery (Table 5)

A.7 Smaller checkpoints

Table 16: Smaller-checkpoint W4A16 quality, action preservation, and task errors. Rates are percentages; ± is three-run SD. $+\tau$ denotes threshold search. MMLU and throughput are point estimates; dashes avoid repeating matched-recall FTR within threshold-only pairs.

Method	FCR ↓	FTR ↓	AbsAcc ↑	MMLU	Tput
Llama-3.1-8B					
FP16 (reference)	—	12.4±0.5	74.6±0.7	67.8	1.00×
GPTQ	28.3±1.4	—	65.2±1.1	66.5	1.91×
AWQ	26.1±1.2	20.3±0.9	66.4±1.0	66.8	1.93×
QuIP#	24.5±1.1	19.1±0.8	67.1±0.9	66.9	1.78×
QAT	22.1±1.0	17.4±0.7	68.3±0.8	67.2	1.92×
GPTQ + τ	20.8±0.9	16.2±0.6	69.1±0.7	66.5	1.91×
FAQ	17.1±0.7	14.1±0.5	71.8±0.6	67.4	1.84×
Qwen-2.5-7B					
FP16 (reference)	—	13.1±0.6	73.2±0.8	71.4	1.00×
GPTQ	29.6±1.5	—	63.8±1.2	70.1	1.88×
AWQ	27.4±1.3	21.2±1.0	64.9±1.1	70.4	1.90×
QuIP#	25.7±1.2	19.8±0.9	65.8±0.9	70.6	1.76×
QAT	23.4±1.0	18.1±0.8	67.0±0.8	70.9	1.89×
GPTQ + τ	22.0±0.9	16.8±0.7	67.8±0.7	70.1	1.88×
FAQ	18.2±0.8	14.6±0.5	70.4±0.6	71.0	1.82×

The 8B and 7B checkpoints test compression sensitivity under the same four-frontier audit. Their MMLU losses and action errors are larger than the corresponding large-model results. FAQ improves both task projections and reference preservation while retaining the same qualitative tradeoff between quality recovery and residual action error.

NeurIPS Paper Checklist

1. Claims

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: Abstract and Sections 3–5 describe the matched-control result, two-controller scope, and measured repair response.

2. Limitations

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: Section 6 states the controller, sampling, task-proxy, and hardware scope and separates preservation from task correctness.

3. Theory assumptions and proofs

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [N/A]

   Justification: The paper reports empirical evaluations and operational metric definitions, not theoretical results requiring proofs.

4. Experimental result reproducibility

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [No]

   Justification: The protocol is specified, but exact audit membership and per-run reconstruction records are not provided.

5. Open access to data and code

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [No]

   Justification: This submission provides the experimental protocol in the paper and technical appendix; it does not provide open access to code or evaluation data.

6. Experimental setting/details

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [No]

   Justification: The paper specifies the split roles, loss parameters, and update budget; a complete optimizer configuration and calibration-bin specification are not provided.

7. Experiment statistical significance

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [No]

   Justification: Run SDs, buffer ranges, and repartition variability are distinguished, but the estimator underlying the paired intervals is not fully specified.

8. Experiments compute resources

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

Answer: [No]

Justification: The GPU configuration, software, and timing workload are specified; per-experiment execution times and total compute are not reported.

9. Code of ethics

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The study evaluates quantization reliability using existing models and conversation benchmarks in accordance with their applicable terms.

10. Broader impacts

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: Section 6 explains how action preservation can reduce unintended changes and why external task tests are needed when the reference policy is erroneous.

11. Safeguards

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: The study evaluates existing models and does not introduce a new high-risk model or dataset.

12. Licenses for existing assets

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [No]

    Justification: Model and benchmark creators are cited, but individual asset licenses and terms are not enumerated.

13. New assets

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [N/A]

    Justification: No new pretrained model or standalone dataset is introduced.

14. Crowdsourcing and research with human subjects

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: The study does not conduct crowdsourcing or human-subject experiments.

15. Institutional review board (IRB) approvals or equivalent for research with human subjects

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: The study does not conduct human-subject experiments.

16. Declaration of LLM usage

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [Yes]

    Justification: The evaluated language-model checkpoints and their role in controller scoring are described in Sections 3 and 5.