ACTION-CONTRACT PRESERVATION: DETECTING AND REPAIRING HIDDEN DECISION DRIFT IN QUANTIZED LLMS

Anonymous authors
Paper under double-blind review

ABSTRACT

A quantized checkpoint enters an action-bearing system through a controller whose score boundaries dispatch tools, invoke retrieval, abstain, or stop generation. Standard benchmark recovery does not determine whether those boundaries still implement the FP16 action policy. We formulate this deployment question as Action-Contract Preservation (ACP): a paired, family-wise test that freezes action semantics, arbitration, operating-point selection, report examples, and repair permissions, and separately checks controller preservation and externally labeled action error. A boundary–perturbation decomposition then isolates two causes of drift—reference requests near a frontier and quantization error larger than their available margin—and motivates BoundaryLock, which combines sensitivity-guided 3/4/8-bit allocation with signed-margin training. Applied to four Llama and Qwen checkpoints, ACP yields explicit qualification decisions. At 90% BFCL recall, it rejects GPTQ with threshold search on Llama-3.1-70B and Qwen-2.5-72B: false tool triggers rise from 8.2% to 11.9$\pm$0.5% and from 7.6% to 10.9$\pm$0.4%, respectively, although MMLU stays within 0.8 points of FP16. The paired audit also detects drift in coupled routing and labeled retrieval and termination endpoints. Against fully crossed QAT with threshold search—matched on data, 500 updates, parameter-weighted bit budget, and threshold rule—BoundaryLock lowers frontier crossing by 1.8–2.3 points and false tool triggers by 0.8–1.7 points across all checkpoints; it retains reported benchmark quality and 1.73–1.84$\times$ FP16 throughput. ACP therefore turns a benchmark-close replacement into an explicit action-contract decision, and its geometry supplies an effective target for low-bit repair.

1 INTRODUCTION

In an action-bearing language-model system, the deployed object is not only a checkpoint. It is the composition of that checkpoint with score normalization, thresholds, arbitration, and named actions: dispatch an API, enter retrieval, abstain, stop, or continue (Schick et al., 2023; Patil et al., 2025; Yao et al., 2023). Replacing FP16 weights with a quantized checkpoint can preserve an answer benchmark while changing this composition at its decision boundaries. A deployment gate based only on recovered MMLU or perplexity therefore evaluates a different observable from the one that determines execution.

This distinction is informational, not merely empirical. Without an assumption coupling benchmark answers to controller margins, identical benchmark outputs place no restriction on paired frontier disagreement (Proposition 3). Existing compression evaluations expose output flips, aggregate agent capabilities, safety degradation, and process errors (Dutta et al., 2024; Dong et al., 2025; Chen et al., 2025; Jang et al., 2026); recent decision-level analyses further characterize correctness agreement and quantization-induced margin shrinkage (Rababah et al., 2026; Wu et al., 2026). These measurements establish that low-bit behavior can move. The unresolved qualification question is whether a particular replacement preserves a pre-existing action interface when its meanings, arbitration, operating-point rule, admissible repair, and report population are held fixed.

Action-Contract Preservation (ACP) makes that interface the evaluand. The contract records the FP16 action semantics and decisions, separates a frozen reference threshold from a frozen rule for

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

calibrating compressed-model thresholds, declares train/calibration/report access and acceptance margins, and evaluates only paired report examples excluded from optimization. ACP retains both controller preservation and external correctness: simultaneous one-sided bounds must satisfy every action-frontier margin and every available labeled-error margin. A single violated bound gives a transparent rejection witness; a pass requires the full conjunction.

The same paired representation yields a repair principle. A compressed decision can cross a reference frontier only when the reference request lies near that frontier or when quantization perturbs its margin beyond the available distance. We formalize this boundary–perturbation decomposition and derive BoundaryLock from its two terms. Sensitivity-guided mixed precision assigns weight bits where block perturbations most affect protected margins, while signed-margin training preserves the reference side of each boundary-critical example. Ordinary QAT receives the same examples, optimizer steps, calibration access, and threshold-search rule, isolating the action-margin objective and allocation.

The empirical study tests this chain on Llama-3.1-8B/70B-Instruct and Qwen-2.5-7B/72B-Instruct. The controlled W4A16 study yields a qualification decision that generic checkpoint summaries cannot supply. At 90% BFCL recall, GPTQ + τ-search raises false tool firing by 3.7 and 3.3 percentage points on the two large models—37 and 33 additional false dispatches per 1,008 negative requests. These runs place MMLU within 0.8 points of FP16, and ARC-Challenge, HellaSwag, ECE, and Brier differ by at most 0.6, 0.6, 0.5, and 0.006 points, respectively. Four independent action frontiers, a coupled five-way router, and labeled retrieval and termination outcomes localize the drift to the declared control interface. BoundaryLock improves FCR by 1.8–2.3 points and tool FTR by 0.8–1.7 points over fully crossed QAT + τ on every checkpoint; all paired BoundaryLock-minus-QAT + τ FCR intervals exclude zero.

ACP supplies a qualification object for quantized action systems: frozen controller semantics, matched repair permissions, paired frontier and labeled-error estimands, and one family-wise decision. The decomposition identifies what to optimize, and the experiments connect its BoundaryLock target to operational errors under matched repairs, component ablations, two controller geometries, supporting bitwidths, and a held-out prompt shift. Together they show how to detect a replacement that aggregate recovery would admit and reduce the responsible drift without trading away the measured efficiency benefit.

2 Related Work and Technical Positioning

Behavioral fidelity under compression. Output-level studies show that similar aggregate accuracy can conceal different per-example answers (Dutta et al., 2024; Rababah et al., 2026). ACBench broadens compression evaluation to workflow, tool-use, retrieval, and application outcomes (Dong et al., 2025); process-level agent analysis shows how a tolerant trajectory score can absorb amplified tool errors (Jang et al., 2026). Wu et al. (2026) model low-bit margin shrinkage and predict binary flips, while Q-resafe (Chen et al., 2025) measures safety degradation and patches safety-critical weights. ACP protects a different object: a named, already deployed controller with fixed action semantics and arbitration. It pairs each report decision to FP16, declares which thresholds and weights may change, combines preservation with external error, and returns an action-wise family decision rather than an aggregate behavioral distance. BoundaryLock complements margin analysis by optimizing the declared controller frontiers under that same held-out decision rule.

Low-bit recovery and mixed precision. GPTQ, AWQ, SmoothQuant, rotations, codebooks, QAT, and low-rank adaptation recover perplexity, benchmark quality, or throughput by improving weight, activation, or output fidelity (Frantar et al., 2023; Lin et al., 2024; Xiao et al., 2023; Dettmers et al., 2022; Yao et al., 2022; Shao et al., 2024; Tseng et al., 2024; Egiazarian et al., 2024; Liu et al., 2024; Li et al., 2024; Dettmers et al., 2023). Sensitivity-guided precision allocation is therefore not by itself our novelty. BoundaryLock differs in the quantity assigned to each block—its perturbation of frozen action margins—and in the signed frontier loss used after allocation. The fully crossed QAT + τ-search comparison gives ordinary QAT the same training data, updates, calibration set, and threshold-repair rule; ranking claims in the primary study are restricted to this matched comparison.

Operating points, selective systems, and tool evaluation. Calibration, selective prediction, and conformal methods estimate probability quality, coverage, or risk while allowing score transforms

or operating-rule selection (Platt, 1999; Guo et al., 2017; Kumar et al., 2019; Mukhoti et al., 2020; Minderer et al., 2021; Geifman & El-Yaniv, 2017; Angelopoulos & Bates, 2021). ACP instead freezes the rule by which a compressed operating point is selected, the FP16 reference decisions, action priority, acceptance margins, and report population. Tool-use benchmarks establish external action correctness (Schick et al., 2023; Patil et al., 2025; Yao et al., 2023; Shinn et al., 2023); ACP asks the complementary replacement question and requires both reference-policy preservation and bounded labeled error. Appendix Table 5 makes these protected and mutable objects explicit.

3 METHOD: ACTION-CONTRACT PRESERVATION

3.1 CONTROLLER SEMANTICS AND FROZEN CONTRACT

Let f_θ be the FP16 model, $\mathcal{A} = \{a_0, \ldots, a_K\}$ an action set with default action a_0, and $s_k(x;\theta)$ the score for protected action a_k. For the independent-margin controller, the four normalized scores are log-probability ratios:

$$s_k(x;\theta) = \log \sum_{v \in V_k} p_\theta(v \mid x) - \log \sum_{v \in V_{\bar{k}}} p_\theta(v \mid x), \tag{1}$$

where V_k and $V_{\bar{k}}$ are disjoint, tokenizer-resolved control-token sets fixed before optimization for tool/continue, retrieve/continue, abstain/answer, and EOS/continue. Log-softmax normalization is computed over each fixed pair. The signed distance and binary decision are

$$d_k(x;\theta) = s_k(x;\theta) - \tau_k, \qquad c_k(x;\theta) = \mathbb{1}[d_k(x;\theta) \geq 0]. \tag{2}$$

For a compressed checkpoint, $\hat{d}_k(x) = s_k(x;\hat{\theta}) - \hat{\tau}_k$ and $\hat{c}_k(x) = \mathbb{1}[\hat{d}_k(x) \geq 0]$. Exact ties trigger the protected action. The post-arbitration action is $A_\theta(x) = \pi(c_1, \ldots, c_K)$ under the frozen priority tool $\succ$ retrieve $\succ$ abstain $\succ$ stop $\succ$ continue, with $A_{\hat{\theta}}$ defined from $\hat{c}$. A native tool call is aligned to BFCL when constrained decoding emits a valid function name and schema. The binary frontiers are the protected controller primitives; A_θ is the action that executes, so ACP records both primitive crossings and post-arbitration action changes.

The second controller uses five pre-arbitration action logits $z_a(x;\theta)$ obtained from the same fixed semantic control-token groups before pairwise subtraction. Its executed decision is

$$C_\theta^{\text{joint}}(x) = \arg\max_{a \in \mathcal{A}} \{z_a(x;\theta) - \tau_a\}, \tag{3}$$

with the same deterministic priority resolving exact ties. Here τ_a are additive action offsets rather than independent binary cutoffs. For preservation accounting, the four indicators $\mathbb{1}[C_\theta^{\text{joint}}(x) = a_k]$ yield action-specific paired decisions; a protected-to-protected substitution changes two indicators, whereas a default-to-protected transition changes one. We therefore distinguish their mean indicator disagreement from exact executed-action disagreement.

The contract is

$$\mathcal{C} = (\mathcal{A}, s, \tau^0, \pi, \mathcal{T}, \mathcal{E}, \mathcal{D}^{\text{train}}, \mathcal{D}^{\text{cal}}, \mathcal{D}^{\text{rep}}, \boldsymbol{\delta}, \boldsymbol{\gamma}).$$

Here τ^0 denotes the frozen FP16 deployment threshold or offset, $\mathcal{T}$ is the predeclared rule that maps compressed-model calibration scores to $\hat{\tau}$, and $\mathcal{E}$ specifies externally labeled endpoints and their denominators. Quantized parameters $\hat{\theta}$ and thresholds $\hat{\tau} = \mathcal{T}(\hat{\theta}, \mathcal{D}^{\text{cal}})$ may change only within the declared repair budget. Action meanings, score normalization, FP16 decisions, arbitration, threshold-selection rule, split membership, margins, and report examples remain fixed. This separates a frozen reference operating point from matched post-quantization retuning.

3.2 PRESERVATION, EXTERNAL ERROR, AND ACCEPTANCE

For report example x and frontier k, define $q_k(x) = \mathbb{1}[c_k(x;\theta) \neq \hat{c}_k(x)]$. ACP reports the crossing vector $\mathbf{r} = (r_1, \ldots, r_K)$, micro FCR, and macro FCR:

$$r_k = \frac{1}{n_k} \sum_x q_k(x), \quad \text{FCR}_{\text{micro}} = \frac{\sum_k n_k r_k}{\sum_k n_k}, \quad \text{FCR}_{\text{macro}} = \frac{1}{K} \sum_k r_k. \tag{4}$$

The main independent controller contributes one decision from each frontier per report turn, making the reported micro and macro values equal. ACP can additionally diagnose executed-action disagreement and its transition matrix:

$$r_{\text{exec}} = \frac{1}{N}\sum_x \mathbb{1}[A_\theta(x) \neq A_{\hat{\theta}}(x)], \qquad T_{ab} = \frac{1}{N}\sum_x \mathbb{1}[A_\theta(x) = a, A_{\hat{\theta}}(x) = b].$$

For the coupled router, replace A with C^{joint}; the indicator FCR and r_{exec} need not coincide because a non-default substitution contributes two indicator changes. For an externally labeled endpoint $g \in \mathcal{G}$ with error e_g, ACP reports excess error $\Delta e_g = e_g(\hat{\theta}) - e_g(\theta)$. Primitive preservation, executed-action preservation, and correctness against external labels are thus separate inspectable quantities. The qualification decision uses the declared primitive frontiers and labeled endpoints; r_{exec} and T are optional diagnostics of how crossings reach execution.

Before model comparison, the deployment owner declares preservation margins δ_k and labeled-error margins γ_g. We use $\delta_k = 2$ percentage points and $\gamma_g = 1$ point. Let $\boldsymbol{\psi} = (r_1, \ldots, r_K, \{\Delta e_g\}_{g\in\mathcal{G}})$. One paired hierarchical bootstrap draw resamples seed clusters and then report identities within each selected seed, using the same draw for every coordinate. The 95th percentile of $\max_j(\hat{\psi}_j^* - \hat{\psi}_j)$, truncated below at zero, is the shared max-statistic critical value $c_{.95}$; hence $U_{.95}(\psi_j) = \hat{\psi}_j + c_{.95} \geq \hat{\psi}_j$. The simultaneous family contains the four protected frontiers and every available labeled endpoint for one model–controller–method checkpoint; models, controllers, and methods define separate qualification decisions. The acceptance rule is

$$\text{Pass}_{\text{ACP}} = \bigwedge_{k=1}^{K} \left[U_{.95}(r_k) \leq \delta_k\right] \wedge \bigwedge_{g\in\mathcal{G}} \left[U_{.95}(\Delta e_g) \leq \gamma_g\right], \tag{5}$$

where $\mathcal{G}$ indexes external outcomes rather than reusing the frontier index. Every coordinate must pass. Conversely, any point estimate above its declared margin is a sufficient rejection witness because its upper bound cannot be smaller. Appendix Table 6 reports the minimal sufficient witnesses for every primary W4A16 row; this distinction lets readers audit the reported decision without mistaking an aggregate method comparison for a complete certificate vector.

3.3 Structural properties and drift decomposition

Proposition 1 (independent-frontier coordinate invariance). For a fixed pair of independent-frontier thresholds, let g_k be strictly increasing, transform both FP16 and quantized scores by $s'_k = g_k \circ s_k$, and set $\tau'_k = g_k(\tau_k)$ and $\hat{\tau}'_k = g_k(\hat{\tau}_k)$. Then every binary decision, crossing indicator, and frontier-preservation rate is unchanged.

Proof. Strict increase gives $s_k(x) \geq \tau_k$ exactly when $g_k(s_k(x)) \geq g_k(\tau_k)$. Applying the equivalence to FP16 and quantized scores preserves both decisions and their disagreement. □

The statement applies after calibration when thresholds are transformed with the scores; it does not assert that an arbitrary threshold-search implementation is equivariant. For the coupled argmax controller, a common positive affine transform of all adjusted logits is sufficient to preserve the selected action, whereas action-specific monotone transforms are not.

Proposition 2 (boundary–perturbation decomposition). Define the perturbation $\Delta_k(x) = \hat{d}_k(x) - d_k(x;\theta)$. For every $\epsilon > 0$,

$$r_k \leq \Pr(|d_k| \leq \epsilon) + \Pr(|\Delta_k| > \epsilon)\,. \tag{6}$$

Proof. If $|d_k| > \epsilon$ and $|\Delta_k| \leq \epsilon$, then $\hat{d}_k$ retains the sign of d_k under the tie rule, so the decision cannot cross. Every crossing therefore belongs to at least one event on the right-hand side. Taking expectations gives Eq. 6. □

Proposition 3 (aggregate quality does not identify contract disagreement). On a report set with N controller decisions, and without an assumption coupling benchmark outputs to frontier margins, identical benchmark outputs are compatible with every empirical ACP crossing rate in $\{0, 1/N, \ldots, 1\}$.

Proof. Benchmark answers and controller margins are distinct observables. Holding all benchmark answers fixed leaves the signs of the N quantized controller margins unspecified by benchmark

accuracy alone. For any integer $j \in \{0, \ldots, N\}$, assigning opposite signs to exactly j paired margins produces empirical FCR j/N without changing a benchmark answer. This is an identifiability statement over observable outputs, not a claim that every assignment is reachable by a particular quantizer. □

Propositions 1–3 establish independent-frontier coordinate invariance, expose the boundary and perturbation terms, and show that benchmark accuracy alone cannot determine action preservation. Equation 6 then identifies the repair target: reduce quantization-induced margin error where the reference controller is most exposed.

4 BOUNDARYLOCK: GEOMETRY-GUIDED CONTRACT REPAIR

BoundaryLock operationalizes both terms of Eq. 6. Let b index transformer blocks, n_b be the number of quantized weights in block b, and $q \in \{3, 4, 8\}$ a candidate precision. For each (b, q), calibrated perturbation draws replace only block b by noise matched to its observed q-bit quantization residual. The sensitivity cost h_{bq} is the average increase, over the fixed draws and boundary-critical training examples, in the signed frontier loss defined below. Dynamic programming then solves

$$\min_{\{q_b\}} \sum_b h_{bq_b} \quad \text{s.t.} \quad \sum_b n_b q_b \le B_{\mathrm{w}}, \qquad q_b \in \{3, 4, 8\},$$

where B_{w} is the parameter-count-weighted payload of the named weight regime. Group size, scale representation, and unquantized modules are fixed across the compared methods; the constraint therefore matches weight payload, while measured throughput provides a separate systems check. At W8A8 the constraint saturates at uniform 8-bit weights, so that coordinate tests training and calibration rather than heterogeneous allocation.

Starting from GPTQ or AWQ, a 500-step AdamW straight-through update protects signs on boundary-critical examples by minimizing

$$\mathcal{L} = \mathcal{L}_{\mathrm{LM}} + \lambda \frac{1}{|\mathcal{D}_{\mathrm{bc}}|} \sum_{x \in \mathcal{D}_{\mathrm{bc}}} \sum_k \max\{0, m - d_k(x; \theta) \hat{d}_k(x)\}, \tag{7}$$

with $\lambda = 0.1$, $m = 0.3$, and boundary set $|d_k| < \epsilon$ for $\epsilon = 0.5$. The product is positive exactly when the FP16 and quantized margins have the same sign; its hinge penalizes sign inversions and insufficient signed agreement. Boundary stratification supplies the near-frontier examples, while the reference-margin factor prioritizes examples whose FP16 side is stable enough to define a training target. Sensitivity-guided allocation separately suppresses blocks that contribute most to $|\Delta_k|$.

Ordinary QAT minimizes $\mathcal{L}_{\mathrm{LM}}$ with the same quantization simulation, trainable weight tensors, AdamW updates, examples, and update count. BoundaryLock adds the frontier term, boundary/interior minibatch stratification, and the mixed-precision allocation. Final thresholds are selected on $\mathcal{D}^{\mathrm{cal}}$ by maximizing agreement with frozen FP16 decisions under the same candidate grid, tie rule, and access used by GPTQ + τ-search and QAT + τ-search. Report examples remain untouched until final paired inference.

5 EXPERIMENTAL DESIGN

5.1 MODELS, DATA FLOW, AND METRICS

We evaluate Llama-3.1-8B/70B-Instruct (Touvron et al., 2023; Grattafiori et al., 2024) and Qwen-2.5-7B/72B-Instruct (Yang et al., 2024). W4A16 is the primary fully crossed comparison; W4A8 and W8A8 are supporting coordinates. Group size is 128. Uniform PTQ baselines use the named weight precision, while BoundaryLock matches the parameter-count-weighted W4 payload with a 3/4/8-bit schedule. All training methods receive 500 updates, and the 128-example PTQ layer-calibration set is disjoint from every ACP split.

The independent controller uses 2,048 turn-level prompts serialized in ShareGPT message format and partitioned into 1,024 train, 512 calibration, and 512 report turns. Thus "ShareGPT-style" denotes the message schema rather than a claim about corpus provenance. Each report turn contributes four

frontier decisions: 2,048 in total, with 808 satisfying $|d_k| < 0.5$ and 1,240 outside that band. The coupled router uses 2,304 turns split into 1,152/576/576, yielding 2,304 protected-action indicators on report. Split construction, score extraction, token groups, and action labels are fixed before method comparison.

BFCL supplies 820 required-call and 1,008 no-call report requests (Patil et al., 2025); false-trigger rate (FTR) is the fraction of no-call requests routed to a valid function name and schema. TriviaQA-uncertain supplies abstention labels (Joshi et al., 2017; Kadavath et al., 2022); AbsAcc is binary accuracy of the abstain/answer action. A held-out retrieval-necessity set contains 824 necessary-retrieval and 1,008 direct-answer requests, with false-route rate (FRR) measured on the latter. A turn-completion set contains 906 stop and 930 continue labels; delayed and premature stopping use these respective denominators. Labels and exclusions are resolved before split construction, and all outcome report sets are disjoint from ACP train and calibration partitions. Generic quality uses 5-shot MMLU on 14,042 test items (Hendrycks et al., 2021), ARC-Challenge on 1,172 items (Clark et al., 2018), and HellaSwag on 10,042 items (Zellers et al., 2019).

5.2 Operating points and inference

ACP preservation and externally labeled fixed-recall evaluation use separate, predeclared threshold traces. For ACP, τ^0 is the FP16 controller threshold and every compressed method applies the same rule $\mathcal{T}$: select on ACP-calibration the candidate threshold with maximum agreement to frozen FP16 decisions, with the tie break in Appendix A, then freeze it for ACP-report. For BFCL and retrieval, a disjoint operating-point calibration partition selects the least permissive threshold attaining at least 90% required-action recall for each checkpoint; that threshold is then frozen and FTR or FRR is computed on untouched labeled report requests. Thus tool-frontier FCR measures preservation at the ACP contract threshold, whereas BFCL FTR is a separate external endpoint at a fixed-recall rule. All methods receive the same calibration identities, candidate thresholds, and tie handling; none selects an operating point from report outcomes.

FCR and labeled action outcomes are mean $\pm$ standard deviation over three independently quantized or trained checkpoints; the single frozen FP16 reference is a point estimate. ECE and Brier evaluate sigmoid-transformed margins $\sigma(d_k) = (1 + \exp(-d_k))^{-1}$ against frozen FP16 binary decisions, macro-averaged across frontiers with one fixed set of bins; they measure calibration to the reference controller, not external correctness. The simultaneous ACP bound uses the seed-aware paired max-statistic bootstrap defined in Section 3. Pairwise effect intervals resample report identities within seed and preserve each identity across compared methods. Exact McNemar tests summarize one pooled binary tool contrast and are reported as complementary discordance checks, not as the family-wise ACP test. MMLU and throughput retain the point-estimate convention, while ARC-Challenge and HellaSwag report item-bootstrap 95% intervals. Experiments run on 8$\times$NVIDIA A100-80GB GPUs with CUDA 12.4 and PyTorch 2.3. Throughput uses batch size 32 and sequence length 2,048. Table 1 reports throughput relative to FP16; Table 9 reports absolute tokens/s on this same stack. Appendix Table 7 traces each protected action from its frozen property and strongest comparator to its held-out outcome and operational failure.

6 Results

6.1 Recovered benchmarks can admit contract drift

The large-model rows expose the decision consequence hidden by benchmark recovery. GPTQ + τ retains MMLU within 0.8 points but reaches 11.9% and 10.9% FTR, 3.7 and 3.3 points above FP16—37 and 33 additional false dispatches per 1,008 negative requests. Each labeled-error estimate exceeds $\gamma_g = 1$, so the corresponding simultaneous upper bound necessarily exceeds the margin. Independently, micro FCRs of 16.1% and 15.4% average four nonnegative frontier rates and therefore guarantee that at least one exceeds $\delta_k = 2$. These are sufficient, reader-checkable witnesses that the benchmark-close replacements do not preserve the declared contract. BoundaryLock improves FCR, FTR, and AbsAcc on every checkpoint; Appendix Table 6 applies the same witness logic to all primary compressed rows.

Figure 1 resolves the aggregate result into the two terms of Eq. 6. The left panel uses evaluation strata defined by the frozen reference distance, not the training cutoff as a changing report denominator.

Table 1: Primary W4A16 comparison on the independent-margin controller. Frontier FCR, BFCL FTR, and AbsAcc are percentages (mean ± std, three seeds); FP16, MMLU, and throughput are point estimates. FTR uses the separately calibrated 90%-recall external endpoint. BoundaryLock ranking claims use the fully matched QAT + τ row; PTQ rows provide context.

Model	Method	FCR ↓	Raw FTR ↓	AbsAcc ↑	MMLU	Tput
Llama-3.1-70B	FP16	—	8.2	81.4	79.2	1.00×
	GPTQ	23.7±1.1	16.5±0.8	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	12.8±0.5	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	14.6±0.7	11.2±0.4	77.7±0.5	78.9	1.83×
	BoundaryLock	12.8±0.6	10.3±0.4	78.9±0.4	79.0	1.76×
Llama-3.1-8B	FP16	—	12.4	74.6	67.8	1.00×
	GPTQ	28.3±1.4	21.8±1.0	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×
	QAT + τ	19.3±0.9	15.6±0.6	69.9±0.7	67.2	1.92×
	BoundaryLock	17.1±0.7	14.1±0.5	71.8±0.6	67.4	1.84×
Qwen-2.5-72B	FP16	—	7.6	82.8	82.3	1.00×
	GPTQ	22.1±1.0	14.9±0.7	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	11.8±0.5	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	14.2±0.6	10.2±0.4	79.4±0.4	82.0	1.80×
	BoundaryLock	12.1±0.5	9.4±0.3	80.7±0.4	82.1	1.73×
Qwen-2.5-7B	FP16	—	13.1	73.2	71.4	1.00×
	GPTQ	29.6±1.5	22.7±1.1	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×
	QAT + τ	20.5±0.9	16.3±0.6	68.9±0.7	70.9	1.89×
	BoundaryLock	18.2±0.8	14.6±0.5	70.4±0.6	71.0	1.82×

ACP preserves the identity of threshold-near requests that a benchmark average discards; the right panel identifies blocks where precision can most reduce the perturbation tail under the fixed parameter-weighted payload.

6.2 Geometry predicts the effective repair components

The decomposition predicts a joint effect from margin protection and perturbation control. Table 10 supports that interaction: removing signed-margin training raises FCR from 12.8% to 16.4%, while replacing the allocation with uniform W4 raises it to 14.9%. The latter is close to QAT + τ at 14.6%, showing that frontier training on a uniform allocation does not explain the full method; the 12.8% result emerges when signed margins and heterogeneous precision act together. Boundary stratification and threshold calibration contribute 1.3 and 0.8 additional points. Against QAT + τ with identical train and calibration access, BoundaryLock reduces FCR by 1.8–2.3 points across all checkpoints, and every paired FCR interval in Appendix Table 8 excludes zero.

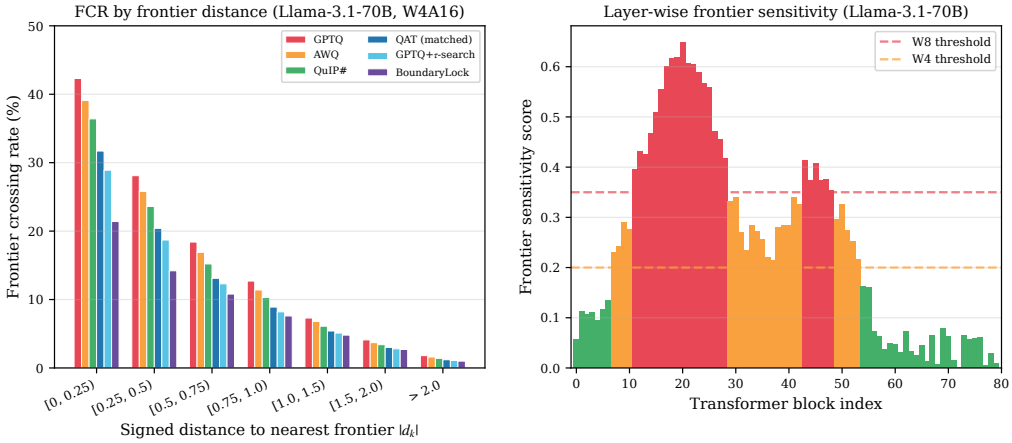

Figure 1: The two terms targeted by BoundaryLock on Llama-3.1-70B W4A16. Left: report-set frontier crossing stratified by frozen FP16 distance $|d_k|$. Right: the normalized block sensitivity cost used by the mixed-precision allocation. Crossings concentrate near a reference boundary, while perturbation sensitivity concentrates in a subset of blocks.

Table 2: Protected-action indicator drift under coupled five-way routing, W4A16, three seeds. A protected-to-protected substitution contributes two indicator changes. Raw/excess FTR is measured against each router's FP16 row; MMLU and throughput reuse Table 1 checkpoints.

Model	Method	FCR ↓	Raw FTR ↓	Excess FTR	MMLU / Tput
Llama-3.1-70B	FP16	—	8.0	—	79.2 / 1.00×
	GPTQ + τ	14.8±0.7	11.4±0.5	3.4	78.4 / 1.82×
	QAT + τ	13.3±0.6	10.7±0.4	2.7	78.9 / 1.83×
	BoundaryLock	11.9±0.5	9.8±0.4	1.8	79.0 / 1.76×
Qwen-2.5-72B	FP16	—	7.3	—	82.3 / 1.00×
	GPTQ + τ	14.1±0.6	10.4±0.4	3.1	81.5 / 1.79×
	QAT + τ	12.8±0.6	9.7±0.4	2.4	82.0 / 1.80×
	BoundaryLock	11.2±0.5	8.9±0.3	1.6	82.1 / 1.73×

6.3 Coupled routing tests controller dependence

Coupling the action logits changes both geometry and accounting but preserves the empirical conclusion. GPTQ + τ remains close on MMLU while exceeding FP16 tool FTR by 3.4 and 3.1 points. BoundaryLock reduces raw FTR relative to GPTQ + τ by 1.6 points (paired 95% CI [0.8,2.4]; discordant false triggers 21/5) and 1.5 points (CI [0.8,2.2]; discordant false triggers 20/5). The result therefore survives a change from independent thresholding to direct five-way competition; the reported FCR remains an action-indicator quantity, while FTR supplies the directly labeled executed-tool endpoint.

6.4 Labeled outcomes test the meaning of a crossing

Table 3: Directly labeled action outcomes on the two large W4A16 models (percent). Compressed rows show mean ± standard deviation over three seeds; FP16 is the fixed reference. Retrieval FRR is measured at 90% necessary-retrieval recall; premature and delayed stop use gold continue and stop denominators.

Model	Method	Retrieval FRR ↓	Premature stop ↓	Delayed stop ↓
Llama-3.1-70B	FP16	7.4	5.9	6.8
	GPTQ + τ	10.5±0.5	8.4±0.4	9.2±0.4
	QAT + τ	9.6±0.4	7.7±0.4	8.5±0.4
	BoundaryLock	8.7±0.4	7.0±0.3	7.8±0.3
Qwen-2.5-72B	FP16	6.9	5.5	6.3
	GPTQ + τ	9.6±0.4	7.8±0.4	8.7±0.4
	QAT + τ	8.9±0.4	7.1±0.3	7.9±0.3
	BoundaryLock	8.1±0.3	6.5±0.3	7.3±0.3

GPTQ + τ increases every labeled retrieval and termination error over FP16. BoundaryLock removes more than half of each excess, preserves the FCR ordering, and has paired improvements over GPTQ + τ whose intervals exclude zero for all six endpoints. FCR therefore aligns with concrete unnecessary retrievals, premature truncations, and delayed terminations rather than merely recording score movement.

6.5 Standard quality and calibration do not reveal the action error

Appendix Table 13 shows ARC-Challenge and HellaSwag within 0.6 points of FP16 and ECE/Brier changes of at most 0.5/0.006, even as GPTQ + τ adds 3.7/3.3 points of false tool firing. Thus generic quality and calibration summaries do not determine the paired decisions measured by ACP, as Proposition 3 predicts.

7 Discussion and Decision Implications

ACP changes the unit of acceptance from a recovered score to a declared action interface. Its paired contract attributes crossings, labeled endpoints connect them to dispatch, routing, and termination errors, and the decomposition identifies where repair can act. BoundaryLock's consistent advantage over fully matched QAT + τ shows the value of optimizing these quantities rather than generic quality alone.

The study covers four action identities, independent and coupled controllers, four Llama/Qwen checkpoints, W4A16 plus supporting bitwidths, labeled outcomes, and a WildChat shift. ACP decisions are conditional on the declared interface, margins, and report population; changing them defines a new evaluation. Within this regime, the reported controllers and endpoints consistently separate aggregate recovery from action preservation, and BoundaryLock ranks above matched QAT + τ on action metrics.

8 Conclusion

Benchmark recovery can mask action-policy changes after quantization. ACP exposes this gap by freezing the contract and jointly bounding paired preservation and labeled error. Across Llama and Qwen checkpoints, it reveals dispatch, retrieval, and termination drift hidden by benchmark and calibration summaries; BoundaryLock improves every checkpoint over matched QAT + τ-search. Low-bit replacements should be qualified by controller decisions, not aggregate recovery.

AI USE STATEMENT

Generative AI tools assisted manuscript preparation, including copyediting, presentation, reference formatting, and submission-readiness review. The authors reviewed the manuscript and retain responsibility for its final content.

REPRODUCIBILITY STATEMENT

Sections 3–5 specify controller computation, data access, admissible retuning, acceptance margins, training budgets, metrics, uncertainty, models, and hardware. Appendix A gives the audit algorithm and threshold-selection rule; the remaining appendices provide paired tests, ablations, held-out shift results, and supporting bitwidth and calibration coordinates.

REFERENCES

Anastasios N Angelopoulos and Stephen Bates. A gentle introduction to conformal prediction and distribution-free uncertainty quantification. *arXiv preprint arXiv:2107.07511*, 2021. URL https://arxiv.org/abs/2107.07511.

Kejia Chen, Jiawen Zhang, Jiacong Hu, Yu Wang, Jian Lou, Zunlei Feng, and Mingli Song. Assessing safety risks and quantization-aware safety patching for quantized large language models. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 9728–9746. PMLR, 2025. URL https://proceedings.mlr.press/v267/chen25ci.html.

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. Think you have solved question answering? try ARC, the AI2 reasoning challenge. *arXiv preprint arXiv:1803.05457*, 2018. URL https://arxiv.org/abs/1803.05457.

Tim Dettmers, Mike Lewis, Younes Belkada, and Luke Zettlemoyer. LLM.int8(): 8-bit matrix multiplication for transformers at scale. *Advances in Neural Information Processing Systems*, 35, 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/c3ba4962c05c49636d4c6206a97e9c8a-Abstract-Conference.html.

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

Peijie Dong, Zhenheng Tang, Xiang Liu, Lujun Li, Xiaowen Chu, and Bo Li. Can compressed LLMs truly act? an empirical evaluation of agentic capabilities in LLM compression. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 14169–14202. PMLR, 2025. URL https://proceedings.mlr.press/v267/dong25k.html.

Abhinav Dutta, Sanjeev Krishnan, Nipun Kwatra, and Ramachandran Ramjee. Accuracy is not all you need. In *Advances in Neural Information Processing Systems*, volume 37, 2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/e0e956681b04ac126679e8c7dd706b2e-Abstract-Conference.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 of *Proceedings of Machine Learning Research*, pp. 12284–12303. PMLR, 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. *Advances in Neural Information Processing Systems*, 30, 2017. URL https://proceedings.neurips.cc/paper/2017/hash/4a8423d5e91fda00bb7e46540e2b0cf1-Abstract.html.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, et al. The Llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024. URL https://arxiv.org/abs/2407.21783.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. On calibration of modern neural networks. *Proceedings of the 34th International Conference on Machine Learning*, 70:1321–1330, 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. *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=d7KBjmI3GmQ.

Jiwon Jang, Kisu Yang, Heuiseok Lim, and Hyunwoo Park. Flat score, amplified failures: How the error budget masks damage in quantized LLM agents. *arXiv preprint arXiv:2607.27275*, 2026. URL https://arxiv.org/abs/2607.27275.

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*, pp. 1601–1611. Association for Computational Linguistics, 2017. doi: 10.18653/v1/P17-1147. URL https://aclanthology.org/P17-1147/.

Saurav Kadavath, Tom Conerly, Amanda Askell, Tom Henighan, Dawn Drain, Ethan Perez, Nicholas Schiefer, Zac Hatfield-Dodds, Nova DasSarma, Eli Tran-Johnson, et al. Language models (mostly) know what they know. *arXiv preprint arXiv:2207.05221*, 2022. URL https://arxiv.org/abs/2207.05221.

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

Yixiao Li, Yifan Yu, Chen Liang, Nikos Karampatziakis, Pengcheng He, Weizhu Chen, and Tuo Zhao. LoftQ: LoRA-fine-tuning-aware quantization for large language models. *International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=LzPWWPAdY4.

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. *Proceedings of Machine Learning and Systems*, 6, 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. *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 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. *Advances in Neural Information Processing Systems*, 34, 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. *Advances in Neural Information Processing Systems*, 33, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/aeb7b30ef1d024a76f21a1d40e30c302-Abstract.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 of *Proceedings of Machine Learning Research*, pp. 48371–48392. PMLR, 2025. URL https://proceedings.mlr.press/v267/patil25a.html.

John C Platt. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. In Alexander J. Smola, Peter L. Bartlett, Bernhard Schölkopf, and Dale Schuurmans (eds.), *Advances in Large Margin Classifiers*, pp. 61–74. MIT Press, 1999.

Baha Rababah, Shahzeb Qamar, Lorenz Sparrenberg, Rafet Sifa, Murat Kantarcioglu, Cuneyt Gurcan Akcora, and Carson K. Leung. The illusion of equivalency: Statistical characterization of quantization effects in LLMs. *arXiv preprint arXiv:2607.08734*, 2026. URL https://arxiv.org/abs/2607.08734.

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. *Advances in Neural Information Processing Systems*, 36, 2023. URL https://openreview.net/forum?id=Yacmpz84TH.

Wenqi Shao, Mengzhao Chen, Zhaoyang Zhang, Peng Xu, Lirui Zhao, Zhiqian Li, Kaipeng Zhang, Gao Peng, Yu Qiao, and Ping Luo. Omniquant: Omnidirectionally calibrated quantization for large language models. *International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=8Wuvhh0LYW.

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

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. LLaMA: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023. URL https://arxiv.org/abs/2302.13971.

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

Zekun Wu, Swati Dhiman, and Adriano Koshiyama. Which decisions low-bit quantization breaks, and how to predict them. *arXiv preprint arXiv:2608.06564*, 2026. URL https://arxiv.org/abs/2608.06564.

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

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, et al. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*, 2024. URL https://arxiv.org/abs/2412.15115.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. ReAct: Synergizing reasoning and acting in language models. *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. *Advances in Neural Information Processing Systems*, 35, 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/adf7fa39d65e2983d724ff7da57f00ac-Abstract-Conference.html.

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

Wenting Zhao, Xiang Ren, Jack Hessel, Claire Cardie, Yejin Choi, and Yuntian Deng. WildChat: 1m ChatGPT interaction logs in the wild. *International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=Bl8u7ZRlbM.

A ACP AND BOUNDARYLOCK DECISION PROTOCOL

Algorithm 1 ACP qualification with BoundaryLock contract-margin repair

Require: FP16 model f_θ, action semantics, τ^0, rule $\mathcal{T}$, margins, target weight payload, frozen train/calibration/report partitions

Cache FP16 scores, frontier decisions, executed actions, and external labels
Estimate h_{bq} on ACP-train by single-block calibrated residual injection
Solve the 3/4/8-bit allocation under $\sum_b n_b q_b \le B_{\mathrm{w}}$
Initialize quantized weights with GPTQ or AWQ
for $t = 1, \ldots, 500$ **do**
    Update quantized weights on ACP-train with Eq. 7
end for
Apply $\mathcal{T}$ on ACP-calibration with the shared grid and tie rule
Run paired FP16 and quantized inference once on untouched ACP-report
Compute the frontier vector, external excess errors, and max-statistic bounds; optionally report execution transitions
Accept exactly when every frontier and labeled endpoint satisfies Eq. 5
return Quantized checkpoint, calibrated thresholds, and ACP decision

For each independent frontier, the ACP threshold grid contains every midpoint between adjacent distinct calibration scores plus the two exterior operating points. Ties are resolved toward the protected action, matching Eq. 2; equal-agreement thresholds are resolved by the smallest absolute displacement from τ^0 and then by the lower threshold. For the joint router, coordinate descent sweeps additive action offsets in fixed tool/retrieve/abstain/stop order until calibration agreement stops improving. The same deterministic rule is applied to GPTQ + τ, QAT + τ, and BoundaryLock. External BFCL and retrieval operating points use their own disjoint calibration partitions and select the least permissive candidate attaining the declared 90% recall before report inference.

Table 4: Protected controller objects and external outcomes. Token groups are tokenizer-resolved once per model and remain fixed across FP16 and compressed checkpoints.

Action	Independent score pair	Executed effect	External report outcome
Tool	tool / continue	dispatch valid function	BFCL FTR on 1,008 no-call requests
Retrieve	retrieve / continue	enter retrieval	FRR on 1,008 direct-answer requests
Abstain	abstain / answer	withhold answer	AbsAcc against answerability label
Stop	EOS / continue	terminate generation	premature on 930 continue; delayed on 906 stop

B Positioning and Evidence Trace

Table 5: Technical positioning of ACP. "Mutable object" records what a method may alter before held-out evaluation.

Line of work	Protected object	Mutable object	Held-out criterion
PTQ/QAT (Frantar et al., 2023; Lin et al., 2024; Liu et al., 2024)	Aggregate LM quality	Weights, scales, bit allocation	Task score or perplexity
Compression agreement (Dutta et al., 2024; Rababah et al., 2026)	Output-level prediction identity	Compression method	Accuracy, agreement, answer flips
Agent-compression evaluation (Dong et al., 2025; Jang et al., 2026)	Agent task or process outcome	Precision, model, prompting	Task score or error channel
Margin-shrinkage analysis (Wu et al., 2026)	Binary decision margins	Quantizer and bitwidth	Flip prediction and margin fit
Quantization safety repair (Chen et al., 2025)	Safety behavior	Safety-critical quantized weights	Safety and utility scores
Calibration/selective prediction (Guo et al., 2017; Geifman & El-Yaniv, 2017)	Probability quality or coverage	Score map, threshold, abstention rule	ECE, risk–coverage, coverage
Tool-routing evaluation (Patil et al., 2025; Schick et al., 2023)	Task and API outcome	Model, prompt, tool policy	Tool accuracy or task success
ACP / BoundaryLock	Frozen action contract and labeled error	Declared repair under a frozen selection rule	Paired bounds, family decision, matched repair

Table 6: Minimal sufficient ACP drift witnesses for every compressed W4A16 independent-controller row. FCR uses 2,048 paired frontier decisions; tool excess uses 1,008 BFCL-negative requests under the separate 90%-recall rule. Micro FCR above 2 points implies at least one frontier-margin violation, and ΔFTR directly exceeds the 1-point labeled-error margin, so each family decision is determined without an aggregate benchmark.

Method	L70 FCR / ΔFTR	L8 FCR / ΔFTR	Q72 FCR / ΔFTR	Q7 FCR / ΔFTR
GPTQ	23.7 / 8.3	28.3 / 9.4	22.1 / 7.3	29.6 / 9.6
AWQ	21.4 / 6.9	26.1 / 7.9	20.3 / 6.2	27.4 / 8.1
QuIP#	19.8 / 6.1	24.5 / 6.7	18.9 / 5.5	25.7 / 6.7
QAT	17.3 / 4.6	22.1 / 5.0	16.7 / 4.2	23.4 / 5.0
GPTQ + τ	16.1 / 3.7	20.8 / 3.8	15.4 / 3.3	22.0 / 3.7
QAT + τ	14.6 / 3.0	19.3 / 3.2	14.2 / 2.6	20.5 / 3.2
BoundaryLock	12.8 / 2.1	17.1 / 1.7	12.1 / 1.8	18.2 / 1.5

Table 6 is a qualification-decision table rather than a method-ranking table. Pairwise method effects use Table 8; ACP drift detection follows from a declared margin lying below an observed endpoint. The coupled-controller rows in Table 2 admit the same direct witness because each excess FTR is above one point.

Table 7: Systems evidence matrix. Quantization is the perturbation; no adaptive attacker is assumed. Every row uses frozen FP16 semantics, matched retuning, and untouched held-out inference.

Protected object	Trust and property	Comparator and held-out test	Outcome, cost, failure, trace
Tool admission	Frozen BFCL mapping; preserve trigger and labeled FTR	GPTQ + τ; BFCL report	Raw/excess FTR; relative throughput; false dispatch; Eq. 5, Tables 1, 2
Retrieval route	Frozen necessity labels; preserve route	QAT + τ; retrieval report	False route; same W4A16 cost; unnecessary retrieval; Table 3
Abstention	Frozen answerability labels; preserve answer/abstain	QAT + τ; TriviaQA-uncertain	AbsAcc; same throughput; incorrect answer policy; Table 1
Stop/continue	Frozen completion labels; preserve termination	QAT + τ; completion report	Premature/delayed stop; same cost; trajectory truncation/extension; Table 3
Coupled router	Frozen joint logits and priority; preserve argmax action	GPTQ + τ; controller-specific report	Joint FCR/FTR; retained MMLU; action substitution; Eq. 3, Table 2

The matrix connects each formal protected object to one observable deployment consequence. In particular, frontier disagreement is always interpreted beside a labeled action outcome and a cost-matched comparator; it is not used as a surrogate for generic task failure.

C Crossed-Control Paired Statistics

Table 8: BoundaryLock minus QAT + τ-search paired effects on the independent controller. Intervals are hierarchical-bootstrap 95% CIs. Discordant tool counts are QAT-only/BoundaryLock-only false triggers; p is the exact McNemar value.

Model	ΔFCR CI	ΔFTR CI	ΔAbsAcc CI	Discordant	p
Llama-3.1-70B	[-2.5,-1.1]	[-1.5,-0.3]	[0.5,1.9]	13/4	0.049
Llama-3.1-8B	[-3.1,-1.3]	[-2.3,-0.7]	[0.8,3.0]	20/5	0.004
Qwen-2.5-72B	[-2.8,-1.4]	[-1.3,-0.3]	[0.6,2.0]	12/4	0.077
Qwen-2.5-7B	[-3.2,-1.4]	[-2.5,-0.9]	[0.4,2.6]	22/5	0.002

The hierarchical intervals estimate the mean paired effect across the three independently quantized or trained checkpoints. McNemar's test instead conditions on the pooled discordant binary tool outcomes, so its finite-count exact p-value need not agree with a seed-aware bootstrap interval. The Qwen-2.5-72B row illustrates this distinction: its bootstrap interval excludes zero while the complementary exact test gives $p = 0.077$; ACP itself uses neither quantity, but the simultaneous one-sided construction in Eq. 5.

Table 9: Access and cost audit on Llama-3.1-70B W4A16. Throughput is absolute tokens/s on 8×A100-80GB, unlike the relative ratios in Table 1.

Method	Updates	Action-interface access	Tput	FCR ↓	MMLU
GPTQ	0	PTQ calibration	4,218	23.7	78.4
AWQ	0	PTQ calibration	4,264	21.4	78.6
GPTQ + τ	Grid	PTQ calibration + ACP-calibration	4,218	16.1	78.4
QAT	500	ACP-train	4,241	17.3	78.9
QAT + τ	500 + grid	ACP-train + ACP-calibration	4,241	14.6	78.9
BoundaryLock	500 + grid	ACP-train + ACP-calibration	4,078	12.8	79.0

The access table isolates the score-bearing method comparison: QAT + τ and BoundaryLock use the same ACP-train examples, 500 updates, ACP-calibration identities, and threshold grid. Their 163-token/s throughput difference is reported rather than hidden, while both retain a substantial advantage over FP16.

D BoundaryLock Ablation

Table 10: BoundaryLock ablation on Llama-3.1-70B W4A16. Each row removes one component; ΔFCR is relative to the full method.

Variant	FCR ↓	ΔFCR	FTR ↓	MMLU
Full BoundaryLock	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 stratification	14.1±0.7	+1.3	11.0±0.4	78.9
− Post-quantization τ calibration	13.6±0.6	+0.8	10.8±0.4	79.0
− All components (plain QAT)	17.3±0.8	+4.5	12.8±0.5	78.9

The signed-margin objective contributes the largest individual gain, while mixed precision and held-out threshold calibration address distinct residual errors. The crossed control in Table 8 confirms that BoundaryLock retains its advantage when ordinary QAT receives identical threshold privileges. For the labeled outcomes in Table 3, paired BoundaryLock-minus-GPTQ + τ 95% intervals are [-2.6,-1.0], [-2.1,-0.7], and [-2.1,-0.7] for Llama-3.1-70B; the corresponding Qwen-2.5-72B intervals are [-2.2,-0.8], [-1.9,-0.7], and [-2.0,-0.8].

E Held-Out Prompt Shift

Table 11: Held-out prompt shift on Llama-3.1-70B W4A16. Controller definitions and thresholds are frozen from the in-distribution ACP protocol.

Method	FCR in-distribution ↓	FCR WildChat ↓	Change
GPTQ	23.7±1.1	26.4±1.3	+2.7
AWQ	21.4±0.9	24.1±1.1	+2.7
GPTQ + τ	16.1±0.7	20.3±0.9	+4.2
BoundaryLock	12.8±0.6	15.6±0.8	+2.8

The thresholds and controller semantics are unchanged under the WildChat (Zhao et al., 2024) prompt shift. BoundaryLock retains a 4.7-point FCR advantage over GPTQ + τ (15.6% versus 20.3%), while its shift-induced increase is 1.4 points smaller. This held-out result tests the same action interface under changed requests rather than selecting a new operating point.

F Supporting Bitwidth Results

The supporting coordinates test whether BoundaryLock's improvement over its quantized initialization persists beyond W4A16. W4A8 retains the same 4-bit weight-payload constraint with quantized activations. At W8A8, the 3/4/8-bit allocation saturates at uniform W8, so the difference reflects signed-margin training and threshold calibration rather than sensitivity-guided bit redistribution. The fully crossed comparator and equal-threshold-privilege analysis remains anchored to W4A16.

Table 12: Supporting bitwidth coordinates on the large checkpoints. FCR and throughput cells report comparator/BoundaryLock point estimates; throughput is relative to FP16. GPTQ + τ is matched at W4A16; GPTQ is the initialization at the supporting coordinates, and W8 uses uniform 8-bit weights.

Model	Bits	Comparator	FCR ↓	ΔFCR	Tput / FP16
Llama-70B	W4A16	GPTQ + τ	16.1 / 12.8	−3.3	1.82 / 1.76×
Llama-70B	W4A8	GPTQ	27.1 / 14.7	−12.4	1.95 / 1.87×
Llama-70B	W8A8	GPTQ	10.4 / 8.7	−1.7	1.42 / 1.39×
Qwen-72B	W4A8	GPTQ	25.8 / 13.9	−11.9	1.92 / 1.84×
Qwen-72B	W8A8	GPTQ	9.8 / 7.1	−2.7	1.39 / 1.36×

G Supporting Quality and Calibration

Table 13: Generic quality and calibration on the same large-model W4A16 runs. ARC-C and HellaSwag show accuracy with item-bootstrap 95% intervals. Compressed-model ECE and Brier show mean ± seed standard deviation; the frozen FP16 reference is a point estimate.

Model	Method	ARC-C ↑	Hella ↑	ECE ↓	Brier ↓	Median $\lvert d \rvert$ [IQR]
L70	FP16	71.8±0.7	87.4±0.4	3.6	0.112	1.18 [0.42,2.31]
	GPTQ + τ	71.2±0.7	86.9±0.4	4.1±0.2	0.118±0.003	1.11 [0.37,2.24]
	QAT + τ	71.4±0.7	87.0±0.4	3.9±0.2	0.116±0.003	1.14 [0.39,2.28]
	BoundaryLock	71.6±0.7	87.2±0.4	3.7±0.2	0.113±0.003	1.16 [0.41,2.30]
Q72	FP16	74.9±0.7	88.6±0.4	3.2	0.104	1.24 [0.46,2.39]
	GPTQ + τ	74.3±0.7	88.1±0.4	3.7±0.2	0.110±0.003	1.17 [0.40,2.31]
	QAT + τ	74.5±0.7	88.2±0.4	3.5±0.2	0.108±0.003	1.20 [0.43,2.35]
	BoundaryLock	74.7±0.7	88.4±0.4	3.3±0.2	0.105±0.003	1.22 [0.45,2.37]