REGAIN: RESTORING TOKEN DISTRIBUTIONS IN LOW-BIT LANGUAGE MODELS VIA TRUNCATED-SUPPORT DISTILLATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Low-bit weight quantization substantially mitigates the memory footprint and latency bottlenecks of large language models, yet aggressive compression regimes incur non-negligible capability degradation. While low-rank quantization error compensation effectively recovers coarse benchmark accuracy with minimal parameter overhead, it fails to restore the nuanced output distributions of the full-precision teacher. Specifically, even when Top-1 predictions appear intact, quantization systematically distorts the probability margins and relative rankings among competing tokens, precipitating severe compounding errors in sequential and structured generation tasks. To resolve this discrepancy, we propose REGAIN, a parameter-efficient recovery framework based on truncated-support distillation. Instead of optimizing over the entire vocabulary or naively mirroring teacher targets, REGAIN dynamically constructs a compact token support from the union of the Top-k candidates from both the full-precision teacher and the quantized student. This design preserves teacher-favored modes while directly penalizing quantization-induced spurious competitors, using an auxiliary Ghost token to aggregate the residual probability mass for bounded computation. With the low-bit backbone frozen, REGAIN exclusively trains low-rank adapters under this compact distributional objective. Extensive evaluations across Qwen3 architectures from 8B to 32B, multiple post-training quantization schemes including GPTQ and QuIP#, and aggressive 2 to 4-bit regimes demonstrate that REGAIN substantially improves distributional fidelity and downstream performance across question answering, tool invocation, and structured reasoning, setting a new Pareto frontier for low-rank recovery of ultra-low-bit language models.

1 INTRODUCTION

Large language models (LLMs) have demonstrated remarkable capabilities in complex reasoning Wei et al. (2022), code generation Yang et al. (2025), and interactive agent workflows Yao et al. (2023a). However, their practical deployment remains constrained by the memory footprint and memory-bandwidth demands of autoregressive inference. Post-training quantization (PTQ) has therefore become a practical paradigm for compressing model weights and reducing memory traffic without the cost of full-model retraining Frantar et al. (2023); Lin et al. (2026); Chee et al. (2023); Tseng et al. (2024a;b); Egiazarian et al. (2024).

Observation. At ultra-low precision, aggressive PTQ inevitably introduces quantization errors, leading to substantial degradation in downstream task performance. Low-rank quantization error compensation (LQEC) mitigates this degradation by introducing a small number of trainable low-rank parameters that compensate for inter-layer quantization errors while retaining the low-bit backbone. We observe that existing LQEC methods recover accuracy on standard log-likelihood-based multiple-choice benchmarks, yet still exhibit measurable quality degradation on function-calling and structured decision-making tasks. We attribute this limitation to the nature of multiple-choice evaluation: by selecting only the highest-scoring answer from a fixed candidate set, it can mask residual differences in model behavior.

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.

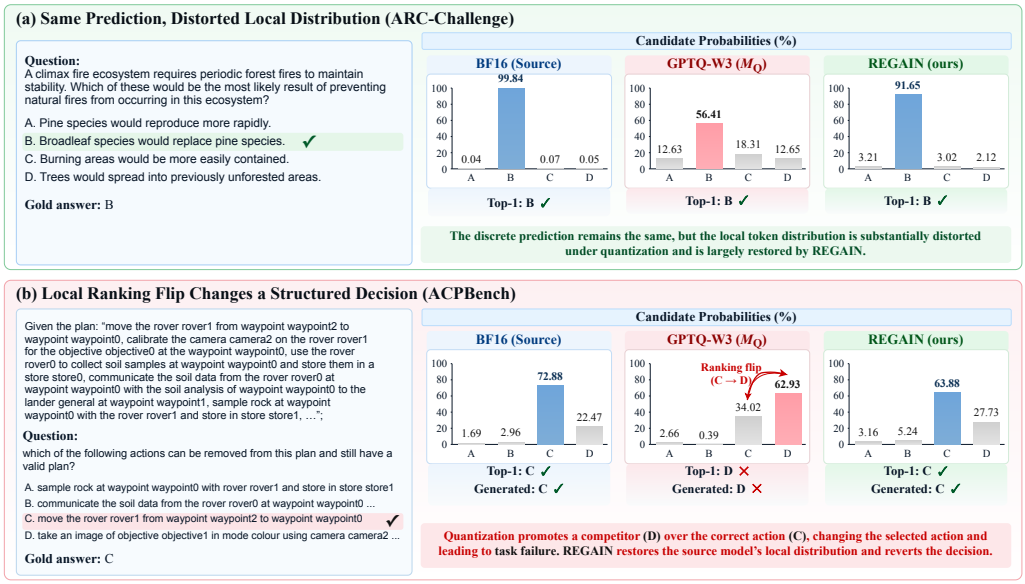

Figure 1: **Low-bit quantization can hide substantial output-preference distortion behind similar discrete predictions.** (a) In multiple-choice evaluation, the top-1 answer can remain unchanged under 2-bit quantization while probability mass shifts toward competing candidates and the teacher-preferred decision margin shrinks substantially. (b) In structured generation, a local ranking reversal at a decision-relevant position can alter the subsequent autoregressive trajectory and result in an invalid function call.

Figure 1 illustrates this mismatch: a fixed-candidate prediction can remain unchanged despite substantial distortion of the local token distribution, whereas a similar local distortion can alter the selected action in structured decision making.

Prediction reversals across different examples can cancel out in aggregate accuracy, while an unchanged top-1 prediction can still exhibit substantial shifts in probability mass among competing candidates. These differences are less visible under fixed-candidate evaluation but become particularly important in autoregressive generation, because changes in token rankings or probabilities directly affect subsequent decoding trajectories. These observations suggest that recovering discrete task predictions alone is insufficient to guarantee recovery of the source model's output behavior, motivating direct constraints on local token probability distributions during low-rank recovery.

Challenge. Full-vocabulary distillation provides dense supervision over token distributions, but capacity-constrained recovery calls for a more targeted and compact objective. Online distillation requires retaining the memory-intensive full-precision teacher throughout adaptation, whereas offline distillation requires caching a vocabulary-sized distribution at every prediction position, incurring substantial storage and I/O overhead. More fundamentally, the limited low-rank adaptation space cannot fit accommodate all vocabulary-level discrepancies. Fine-grained supervision should therefore be allocated to candidates that are most relevant to local decisions.

Teacher-only Top-k distillation ignores model-specific errors exposed by the quantized model itself. Moreover, simply truncating and renormalizing the teacher distribution redistributes omitted probability mass to the retained tokens, thereby distorting the original distribution. An effective compact recovery target should therefore preserve teacher-preferred candidates, explicitly expose quantization-induced competitors, and retain the aggregate probability mass outside the explicit support.

Our Proposal. We propose REGAIN, a parameter-efficient recovery framework grounded in *quantization-aware support distillation*. At each token position, REGAIN constructs a compact distillation support by taking the union of top-k prediction candidates from the full-precision teacher

and the quantized student. The teacher identifies dominant reference tokens to preserve, while the quantized student exposes distorted competitor tokens that require explicit probability penalization. To account for probability leakage outside this restricted support without computing full-vocabulary normalizations, we incorporate a virtual *Ghost* class that aggregates the residual probability mass under the teacher distribution. These sparse supervision targets are generated entirely offline, which removes the memory overhead of hosting the teacher model during recovery and avoids caching dense vocabulary-wide logits.

We evaluate REGAIN on Qwen3 models from 8B to 32B parameters under aggressive 2–4-bit weight quantization via GPTQ and QuIP#. Beyond standard zero-shot benchmarks, we evaluate mathematical reasoning on MATH500 Lightman et al. (2024), downstream decision robustness on the Berkeley Function Calling Leaderboard (BFCL V3) Patil et al. (2025), and structured planning on ACPBench Kokel et al. (2025). Across these settings, REGAIN consistently restores source-model fidelity and narrows the accuracy gap left by conventional low-rank error compensation.

Our main contributions are summarized as follows:

- We diagnose that conventional low-rank error compensation methods fail to prevent severe token-level probability distribution distortions under ultra-low bit-widths. We demonstrate that these residual distortions misalign local decision boundaries, substantially degrading performance on decision-critical tasks such as multi-step reasoning and structured generation.
- To address this distributional collapse, we propose REGAIN, a parameter-efficient recovery framework grounded in quantization-aware support distillation. REGAIN explicitly aligns predictions over a concentrated candidate support by jointly supervising teacher-preferred tokens and quantization-induced competitor tokens, while introducing a compact Ghost class to regularize the aggregate residual tail probability.
- Across diverse model architectures, quantization algorithms, and extreme bit-widths (e.g., sub-4-bit regimes), REGAIN consistently preserves distributional fidelity to full-precision teachers. Consequently, our approach establishes new Pareto frontiers across demanding benchmarks, spanning complex reasoning, structured function calling, and embodied action planning.

2 RELATED WORK

2.1 ULTRA-LOW-BIT LLM QUANTIZATION

PTQ has progressed from second-order and activation-aware scalar quantization Frantar et al. (2023); Lin et al. (2026) to more structured ultra-low-bit methods based on incoherence processing, lattice or trellis codebooks, and additive quantization Chee et al. (2023); Tseng et al. (2024a;b); Egiazarian et al. (2024). These methods substantially reduce the gap between low-bit and full-precision models on perplexity and standard downstream benchmarks. However, strong aggregate performance does not guarantee that complex generative behavior is equally preserved. Li et al. Li et al. (2026) show pronounced degradation in mathematical reasoning under aggressive quantization, while LFQ Lee et al. (2026) reports larger losses on complex generation than on conventional language-understanding tasks. Rather than redesigning the quantizer, we focus on recovering behavioral discrepancies that remain after quantization.

2.2 LOW-RANK QUANTIZATION ERROR COMPENSATION

LQEC augments a frozen quantized backbone with a small number of high-precision low-rank parameters. Existing methods recover quantization error at different levels: ZeroQuant-V2 Yao et al. (2023b) and LQER Zhang et al. (2024) reconstruct weight-level residuals; LoftQ Li et al. (2024) jointly optimizes quantization and low-rank initialization; QERA Zhang et al. (2025) incorporates activation-aware reconstruction; and RILQ Lee et al. (2025) coordinates recovery through model-level representation matching. Recover-LoRA Das et al. (2025) further introduces logit distillation, showing that output-space supervision can benefit low-rank recovery. REGAIN studies a complementary question: under a constrained low-rank recovery budget, which output relationships should receive fine-grained supervision?

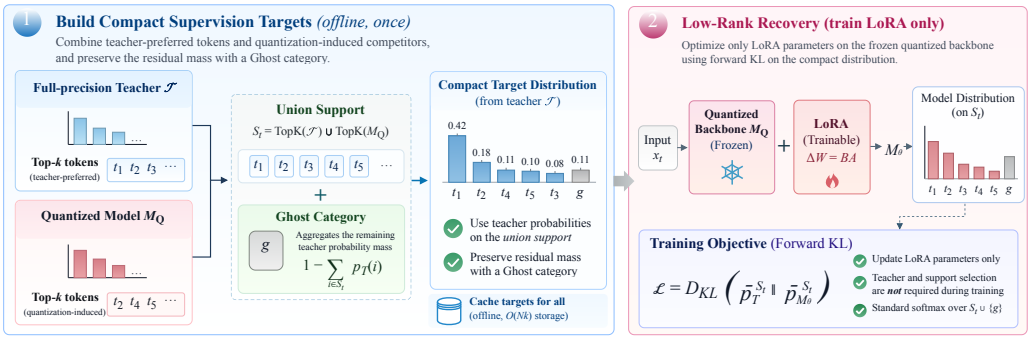

Figure 2: **Overview of REGAIN.** The teacher and the initial quantized model jointly define a compact, quantization-aware support. Teacher probabilities on this support, together with a Ghost residual category, form the offline supervision targets for LoRA-only recovery of the frozen quantized backbone.

2.3 OUTPUT-AWARE AND SPARSE DISTRIBUTION DISTILLATION

Full-Precision vs. Output-Level Alignment. While traditional post-training quantization primarily minimizes layer-wise intermediate reconstruction errors, LFQ Lee et al. (2026) demonstrates the necessity of directly aligning final output distributions. However, LFQ optimizes full-precision weights during block-wise calibration, whereas we target post-quantization recovery over a frozen low-bit backbone via parameter-efficient adaptors.

Sparse Distillation and Our Formulation. Supervised fine-tuning over full-vocabulary logits incurs prohibitive offline storage overhead. To mitigate this, Sparse Logit Sampling Anshumann et al. (2025) addresses top-k truncation distortions by introducing a Ghost token to aggregate the residual probability mass. Despite preserving tail mass, such generic distillation strategies remain quantization-agnostic. They fail to identify the spurious competitor tokens introduced by aggressive weight discretization. REGAIN addresses this limitation by tailoring the distillation support to the specific error profile of the quantized student. While building upon residual-mass preservation principles, our approach explicitly uncovers and penalizes distorted competitor tokens, enabling targeted probability correction within the capacity constraints of low-rank adaptation.

3 REGAIN

We present REGAIN, a post-quantization recovery framework that restores decision-level fidelity compromised by aggressive weight quantization. Instead of optimizing over the entire vocabulary, REGAIN constructs a compact, position-specific recovery target. At each token position, we extract a sparse candidate set combining top predictions from both the full-precision teacher and the quantized model, while aggregating the remaining probability mass into a single residual category. These targets are computed entirely offline and supervise the low-rank adaptation parameters while keeping the low-bit model backbone frozen. Figure 2 illustrates the overall pipeline.

3.1 PROBLEM FORMULATION

We formulate post-quantization recovery as a parameter-efficient distribution alignment problem over a frozen, aggressively quantized backbone. Let M_T denote the BF16 teacher model, M_Q its initial weight-quantized counterpart, and M_θ the recovered model obtained by augmenting M_Q with trainable low-rank parameters θ. For each linear layer parameterized by quantized weights $\hat{W}$, we augment it with low-rank adaptation matrices $\Delta W_\theta = BA$, where only the low-rank parameters $\theta = \{A, B\}$ are trainable while $\hat{W}$ remains frozen. Given a recovery corpus $\mathcal{D}$, all models process identical prompt prefixes under teacher forcing. Specifically, for an input sequence $x = (x_1, \ldots, x_L) \in \mathcal{D}$ and each token index $t \in [1, L]$, let $c_t = x_{<t}$ represent the ground-truth prefix. The predictive distributions over the full vocabulary $\mathcal{V}$ produced by the teacher, the initial

quantized student, and the adapted student are respectively denoted as:

$$p_{M_T}(\cdot \mid c_t), \qquad p_{M_Q}(\cdot \mid c_t), \qquad p_{M_\theta}(\cdot \mid c_t), \tag{1}$$

Conventional full-distribution distillation optimizes the Kullback-Leibler divergence between p_T and p_{M_θ} over all $v \in \mathcal{V}$, which incurs severe compute and memory bottlenecks during offline caching and online normalization. Instead, our objective is to faithfully recover the *local token distribution* over a sparse, decision-critical candidate set $\mathcal{S}_t \subset \mathcal{V}$ at each step t, aligning local decision margins under the parameter constraints of θ.

3.2 Quantization-Aware Support Construction

For each prediction position with context c_t, REGAIN constructs an asymmetric distillation support by taking the union of the top-k candidates from both the teacher and the initial quantized model:

$$\mathcal{S}_t = \mathrm{TopK}\big(p_{M_T}(\cdot \mid c_t),\, k\big) \;\cup\; \mathrm{TopK}\big(p_{M_Q}(\cdot \mid c_t),\, k\big). \tag{2}$$

By construction, the support size is strictly bounded: $k \le |\mathcal{S}_t| \le 2k \ll |\mathcal{V}|$.

The teacher partition, $\mathrm{TopK}(p_{M_T}, k)$, identifies salient tokens that define the ground-truth decision boundary. Conversely, aggressive weight quantization often disrupts relative logit margins, promoting spurious competitor tokens into the high-confidence regime. The student partition, $\mathrm{TopK}(p_{M_Q}, k)$, explicitly isolates these distorted candidates. Importantly, p_{M_Q} acts solely as an error indicator to select which coordinates require supervision; all corresponding supervisory probabilities are strictly derived from the teacher p_{M_T}.

Because $\mathcal{S}_t$ depends only on the initial, unadapted state M_Q, the support indices and target probabilities are computed once in a single offline forward pass. This static construction avoids online inference of the full-precision teacher and circumvents the prohibitive memory requirement of caching dense vocabulary-wide distributions.

3.3 Residual Probability Preservation via Ghost Class

After identifying the active support $\mathcal{S}_t$, naively truncating and renormalizing the probabilities over $\mathcal{S}_t$ would artificially inflate the probability mass of the selected tokens, shifting prediction entropy and corrupting calibration. To preserve the true magnitude of the teacher's distribution without expanding the label space back to $|\mathcal{V}|$, we introduce a virtual residual category g, following the formulation in Sparse Logit Sampling Anshumann et al. (2025).

For each position t, the compact target distribution over the augmented set $\mathcal{S}_t \cup \{g\}$ is formally defined as:

$$\bar{p}_T^{\mathcal{S}_t}(u \mid c_t) = \begin{cases} p_T(u \mid c_t), & u \in \mathcal{S}_t, \\ 1 - \displaystyle\sum_{v \in \mathcal{S}_t} p_T(v \mid c_t), & u = g. \end{cases} \tag{3}$$

We define the recovered model's distribution $\bar{p}_\theta^{\mathcal{S}_t}$ over $\mathcal{S}_t \cup \{g\}$ analogously.

Under this formulation, every explicit candidate $u \in \mathcal{S}_t$ maintains its uncompressed probability under the original full-vocabulary normalization, while g acts as a sink capturing the aggregate probability of the neglected tail. Consequently, support selection dictates the precise coordinates that receive fine-grained, token-level supervision, whereas the remainder of the vocabulary is regulated through its aggregate mass.

3.4 Distillation Objective and Low-Rank Optimization

Given N valid prediction positions across the recovery corpus, we optimize the forward Kullback-Leibler divergence between the teacher and recovered compact distributions:

$$\mathcal{L}_{\mathrm{REGAIN}} = \frac{1}{N} \sum_{t=1}^{N} D_{\mathrm{KL}} \left(\bar{p}_T^{\mathcal{S}_t} \,\middle\|\, \bar{p}_{M_\theta}^{\mathcal{S}_t} \right). \tag{4}$$

During recovery, the quantized backbone weights remain strictly frozen, and only the low-rank adaptation parameters θ receive gradient updates.

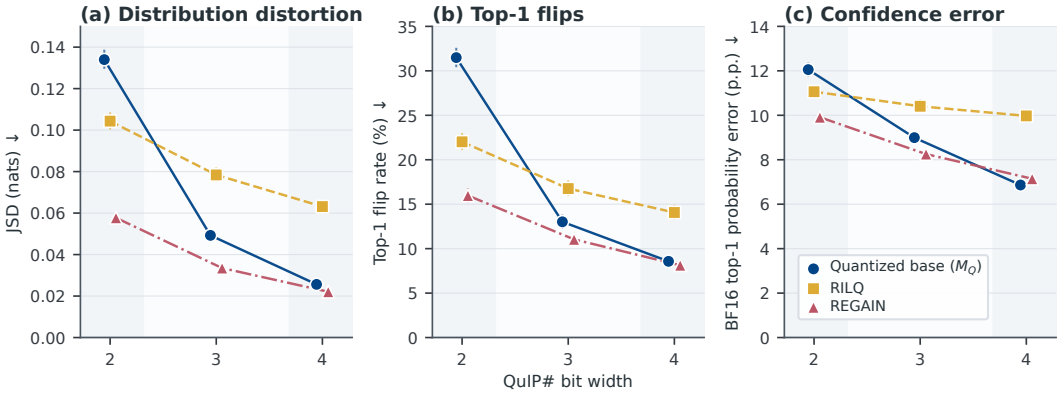

Figure 3: **Residual token-level distortion after low-rank recovery.** M_Q, RILQ, and REGAIN are compared with the BF16 source model under Qwen3-32B QuIP# W2–W4. The panels report Jensen–Shannon divergence, top-1 disagreement, and source top-1 probability error. Lower is better.

Offline Storage and I/O Efficiency. All supervisory targets are generated offline in a single forward pass through each model. For each token position, we store only the selected support indices $\mathcal{S}_t$, their corresponding teacher probabilities $p_T(u \mid c_t)$, and the residual scalar $p_T(g \mid c_t)$. Because $|\mathcal{S}_t| \leq 2k$ with $k \ll |\mathcal{V}|$, caching these targets scales as $O(Nk)$ rather than $O(N|\mathcal{V}|)$. This yields orders-of-magnitude reductions in offline disk storage and data-loading bandwidth during training.

Capacity-Aware Gradient Allocation. Beyond system-level savings, this compact formulation resolves a fundamental optimization bottleneck in low-rank adaptation. Standard full-vocabulary distillation disperses gradient updates across hundreds of thousands of low-probability tail coordinates, which quickly saturates the low-rank subspace of LoRA. By confining token-level supervision strictly to $\mathcal{S}_t$ and aggregating the remaining tail into the single Ghost class, REGAIN concentrates parameter updates exclusively on active decision competitors, matching the fine-grained recovery signal to the constrained capacity of low-rank adaptors.

4 EXPERIMENTS

Models. We evaluate REGAIN on Qwen3-8B, 14B, and 32B Yang et al. (2025). Models are quantized via two standard post-training algorithms: GPTQ Frantar et al. (2023) and QuIP# Tseng et al. (2024a), spanning 2-bit, 3-bit, and 4-bit weight regimes. Qwen3-32B serves as the primary benchmark for extreme low-bit recovery, while the 8B and 14B models evaluate scalability across parameter scales. For each setting, we benchmark the full-precision (BF16) reference M_T, the unrecovered quantized model M_Q, and the adapted student M_θ.

Baselines. We compare against RILQ Lee et al. (2025), the primary low-rank error compensation baseline, alongside the unrecovered baseline M_Q to measure post-quantization degradation. To guarantee a fair comparison, RILQ and REGAIN share the same calibration corpora, training token budgets, LoRA parameterizations, and optimization schedules. In controlled settings, we also benchmark against full-vocabulary forward KL distillation as an idealized upper-bound reference, which becomes computationally intractable for full-scale evaluations.

Implementation Details. For low-rank adaptation, we set the support size parameter to $k = 32$ and initialize LoRA adapters on all linear projection layers. All offline supervision targets are precomputed using the unadapted M_Q and full-precision M_T over a shared calibration corpus of 294K tokens. Evaluation covers two primary regimes: zero-shot multiple-choice accuracy across ARC-Challenge, ARC-Easy, BoolQ, PIQA, and WinoGrande Clark et al. (2018; 2019); Bisk et al. (2020); Sakaguchi et al. (2021), as well as generative execution spanning mathematical reasoning on MATH500 (Avg@8), tool use on BFCL V3 Patil et al. (2025), and structured planning on ACPBench Kokel et al. (2025) using greedy decoding.

Table 1: **Main capability results on Qwen3-32B.** 0-shot Avg. is the macro-average accuracy over ARC-Challenge, ARC-Easy, BoolQ, PIQA, and WinoGrande. MATH500 reports mean accuracy over eight independent stochastic generations per problem (Avg@8), with standard deviation across the eight runs. BFCL V3 and ACPBench use greedy decoding. Within each quantization setting, the best result among M_Q, RILQ, and REGAIN is shown in bold.

Model	Quantizer	Bits	Method	0-shot Avg.↑	MATH500 Avg@8↑	BFCL V3↑	ACPBench↑
Qwen3-32B	–	BF16	Source	76.28	76.37 ± 0.85	83.10	72.61
	GPTQ	2	M_Q	40.59	0.00 ± 0.00	29.59	2.39
			RILQ	65.90	14.89 ± 0.96	29.62	36.28
			REGAIN	**66.46**	**21.49 ± 0.59**	**30.94**	**39.83**
		3	M_Q	71.32	70.83 ± 0.46	74.98	63.78
			RILQ	73.91	71.17 ± 0.54	67.93	66.72
			REGAIN	**74.00**	**72.89 ± 0.89**	**79.21**	**69.00**
		4	M_Q	75.29	**76.40 ± 1.02**	81.87	71.17
			RILQ	74.11	74.80 ± 0.91	78.29	70.78
			REGAIN	**75.92**	76.10 ± 0.53	**82.63**	**72.50**
	QuIP#	2	M_Q	53.96	8.10 ± 0.57	33.24	33.00
			RILQ	65.96	38.30 ± 1.35	46.44	49.39
			REGAIN	**70.96**	**49.75 ± 2.37**	**61.18**	**51.28**
		3	M_Q	72.93	71.85 ± 0.77	77.97	69.67
			RILQ	74.79	70.75 ± 0.62	79.22	66.28
			REGAIN	**75.65**	**73.55 ± 1.19**	**79.44**	**69.72**
		4	M_Q	74.57	74.90 ± 0.52	82.54	**73.06**
			RILQ	75.45	74.40 ± 0.42	80.26	72.83
			REGAIN	**76.08**	**75.40 ± 0.81**	**82.60**	**73.06**

4.1 MAIN RESULTS

Table 1 reports Qwen3-32B results across all bit-widths, while Table 2 details cross-scale evaluations on Qwen3-8B and Qwen3-14B. Across the 32B sweep, REGAIN outperforms RILQ across all six combinations of quantizers (GPTQ, QuIP#) and bit-widths (2, 3, and 4 bits). While improvements are consistent on discriminative zero-shot benchmarks, gains are markedly larger on generative, structured tasks. On Qwen3-32B under 2-bit GPTQ, REGAIN increases MATH500 Avg@8 from 14.89 ± 0.96 to 21.49 ± 0.59, an absolute gain of 6.60 points over RILQ. This disparity is pronounced across scales: for Qwen3-14B with 2-bit QuIP#, REGAIN improves over RILQ by 2.25 points on zero-shot average, but surges by 7.34 points on BFCL V3 and 32.39 points on ACPBench. As bit-width increases, unrecovered models retain higher fidelity to the BF16 teacher, narrowing recovery headroom. Conversely, under aggressive 2-bit regimes, standard low-rank recovery partially restores multiple-choice accuracy but leaves severe regressions on autoregressive decision trajectories. By explicitly penalizing competitor tokens, REGAIN resolves this divergence, restoring generative reliability where feature-level compensation fails.

4.2 RESIDUAL TOKEN-LEVEL DISTORTION

The main results motivate a closer look at the output distributions that remain after low-rank recovery. We analyze Qwen3-32B under QuIP# W2–W4 on shared held-out prediction contexts, comparing M_Q, RILQ, and REGAIN with the BF16 source model. Metric definitions and additional same-top1 analyses are provided in the appendix.

Residual Distortion After Recovery. Figure 3 reports three token-level recovery diagnostics: Jensen–Shannon divergence (JSD) from the full-precision source distribution, Top-1 disagreement with the source model, and absolute error in the source's Top-1 probability. Residual distortion grows sharply as bit width decreases, confirming that ultra-low-bit quantization perturbs both local distributions and token decisions.

Across W2–W4 and all three diagnostics, REGAIN achieves lower error than RILQ in every comparison, roughly halving JSD and reducing both Top-1 disagreement and Top-1 probability error.

Table 2: **Cross-scale capability results.** We report Qwen3-8B and Qwen3-14B results under GPTQ and QuIP#. 0-shot Avg. is the macro-average accuracy over ARC-Challenge, ARC-Easy, BoolQ, PIQA, and WinoGrande. BFCL V3 reports the official overall score, and ACPBench uses greedy decoding. Within each model, quantizer, and bit-width setting, the best result among M_Q, RILQ, and REGAIN is shown in bold.

			0-shot Avg.↑		BFCL V3↑		ACPBench↑	
Model	Quantizer	Method	W2	W3	W2	W3	W2	W3
Qwen3-14B	–	BF16 Source	75.48		83.16		72.50	
	GPTQ	M_Q	42.26	53.89	19.73†	53.89	2.78	59.83
		RILQ	61.06	53.76	19.73†	53.76	**17.06**	57.33
		REGAIN	**62.85**	**54.52**	**20.00**	**54.52**	16.72	**60.61**
	QuIP#	M_Q	47.34	69.18	20.65	57.34	8.50	**60.83**
		RILQ	66.95	72.50	37.70	**58.37**	15.89	57.78
		REGAIN	**69.20**	**74.22**	**45.04**	58.36	**48.28**	59.44
Qwen3-8B	–	BF16 Source	73.28		82.10		68.22	
	GPTQ	M_Q	38.99	67.75	**19.73**†	50.60	1.67	55.11
		RILQ	54.31	**70.87**	**19.73**†	52.92	4.89	51.89
		REGAIN	**55.01**	70.85	**19.73**†	**54.21**	**11.17**	**55.89**
	QuIP#	M_Q	43.34	62.69	19.73†	43.47	4.17	47.61
		RILQ	62.76	69.72	19.73†	52.28	17.06	44.22
		REGAIN	**62.91**	**70.95**	**31.36**	**53.23**	**22.61**	**48.44**

† Degenerate-generation case. The model produces no valid function calls and obtains its non-zero BFCL score primarily from irrelevance-detection cases, where no function call is the expected output.

These results show that REGAIN does not merely improve downstream scores—it restores the source model's local token-level behavior more accurately than conventional low-rank recovery.

Crucially, this advantage does not stem solely from correcting Top-1 prediction flips: even at positions where the recovered model matches the source Top-1, substantial distributional and confidence discrepancies remain, meaning identical discrete predictions can conceal meaningful local distribution differences. We report the full same-Top-1 analysis in the appendix B.

Quantization-induced Competitors. We additionally observe that quantization introduces candidates that are absent from the source-model Top-k but enter the Top-k of M_Q. For $k = 32$, these candidates constitute 56.6%, 25.2%, and 13.0% of the M_Q Top-k under GPTQ W2, W3, and W4, respectively. Thus, teacher-only Top-k supervision can omit candidates that become locally competitive after quantization. This observation motivates the union support evaluated in Section 4.4.

4.3 SUPERVISION EFFICIENCY

Dense output distillation admits two common implementations. Online Full-KL avoids storing teacher targets but requires the BF16 teacher throughout recovery. Offline Full-KL removes repeated teacher inference by caching a full-vocabulary target at every prediction position. REGAIN follows the offline setting but stores only the quantization-aware support and residual mass.

Table 3 summarizes the resulting trade-off under the same controlled recovery setup. Offline Full-KL requires 89.44 GB of cached BF16 full-vocabulary log-probabilities, whereas the REGAIN cache occupies only 247 MB, a 362× reduction.

The runtime improvement is intentionally more modest than the storage reduction. Both approaches still compute full-vocabulary logits during target construction; REGAIN primarily compresses the supervision representation rather than eliminating full-vocabulary output computation.

Table 3: **Supervision efficiency on the controlled Qwen3-8B setting.** Offline Full-KL caches BF16 full-vocabulary log-probabilities; REGAIN stores the Union-32 compact target. Training throughput is measured under matched recovery settings.

Supervision	Teacher in training	Cache	Train tok/s↑
Online Full-KL	Yes	–	1,607
Offline Full-KL	No	89.44 GB	1,559
REGAIN	No	**247 MB**	**1,913**

Table 4: **Component analysis on Qwen3-8B with 2-bit QuIP#.** Results are mean±std over four seeds. 0-shot Avg. is the five-task macro-average; Agree. denotes top-1 token agreement with the BF16 source model, and Cand. JSD measures candidate-distribution divergence.

Support	Target / Residual	0-shot Avg.↑	Agree.↑	Cand. JSD↓
Support construction				
Teacher Top-32	Ghost	63.03±0.14	78.73±0.51	0.0903±0.0019
M_Q Top-32	Ghost	60.62±0.17	76.85±0.20	0.1009±0.0006
Teacher Top-m_t	Ghost	62.91±0.66	79.05±0.29	0.0893±0.0010
Teacher Top-32 + random	Ghost	63.03±0.47	79.05±0.50	0.0894±0.0018
Union Top-32	Ghost	**63.10±0.19**	**79.28±0.12**	**0.0883±0.0004**
Residual-mass treatment				
Union Top-32	Conditional KL	62.09±0.37	77.77±0.20	0.0929±0.0007
Union Top-32	Mass-weighted tail	61.71±0.17	77.75±0.12	0.0927±0.0003
Union Top-32	Ghost	**63.10±0.19**	**79.28±0.12**	**0.0883±0.0004**
Dense-supervision reference				
Full vocabulary	Forward KL	62.91±0.33	79.04±0.26	0.0886±0.0008

4.4 ABLATION AND SENSITIVITY TO SUPPORT BUDGET

Component Ablation. Table 4 decouples candidate support construction from residual probability preservation. Using only the M_Q Top-32 performs markedly worse than teacher-based support, showing that student candidates must complement rather than replace ground-truth teacher preferences. Under a strictly matched budget ($|\mathcal{S}_t| = 32$), the union support achieves the best zero-shot accuracy, teacher agreement, and candidate-level JSD. These controlled comparisons verify that performance improvements stem from targeted competitor selection rather than supervising more tokens. Furthermore, replacing the Ghost category with either conditional KL or a mass-weighted tail consistently degrades accuracy and teacher fidelity, highlighting the necessity of an explicit residual sink. Finally, compact REGAIN matches or slightly exceeds the dense full-vocabulary KL baseline while drastically reducing supervisory footprint.

Support-budget Sensitivity. We vary the per-branch Top-k budget over $k \in \{8, 16, 32, 64\}$ under W2 and W3. Under W2, 0-shot Avg. improves from 58.59 at $k = 8$ to 62.11 at $k = 64$, with most of the gain already realized by $k = 32$ (61.68). Under W3, performance remains nearly unchanged across the tested range, from 70.47 to 70.58. We therefore use $k = 32$ as the default, which provides a compact support while avoiding reliance on a narrowly tuned budget.

5 CONCLUSION

We presented REGAIN, a parameter-efficient recovery framework that restores token-level decision fidelity in aggressively quantized LLMs. By constructing an asymmetric support spanning both teacher preferences and quantization-induced competitor tokens, and regularizing the remaining probability mass via a virtual Ghost class, REGAIN aligns critical decision boundaries without incurring dense full-vocabulary overhead. Across Qwen3 models from 8B to 32B under 2-to-4-bit GPTQ and QuIP# regimes, REGAIN consistently outperforms conventional low-rank error compensation, matching or exceeding dense full-vocabulary distillation while drastically reducing the offline supervisory footprint.

AI USE STATEMENT

Generative AI tools were used to provide feedback on research methodology and experimental design, assist with method implementation and translation, and support the interpretation of experimental results. All AI-assisted work was reviewed and verified by the authors. The remaining required-disclosure activities were either not performed with generative AI or were not applicable to this work. The authors take responsibility for the final content of this work, including claims and artifacts produced with the aid of generative AI.

REFERENCES

Anshumann, Mohd Abbas Zaidi, Akhil Kedia, Jinwoo Ahn, Taehwak Kwon, Kangwook Lee, Haejun Lee, and Joohyung Lee. Sparse logit sampling: Accelerating knowledge distillation in LLMs. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 18085–18108, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-251-0. doi: 10.18653/v1/2025.acl-long.885. URL https://aclanthology.org/2025.acl-long.885/.

Yonatan Bisk, Rowan Zellers, Jianfeng Gao, Yejin Choi, et al. Piqa: Reasoning about physical commonsense in natural language. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pp. 7432–7439, 2020.

Jerry Chee, Yaohui Cai, Volodymyr Kuleshov, and Christopher De Sa. QuIP: 2-bit quantization of large language models with guarantees. In *Advances in Neural Information Processing Systems*, 2023. URL https://arxiv.org/abs/2307.13304.

Christopher Clark, Kenton Lee, Ming-Wei Chang, Tom Kwiatkowski, Michael Collins, and Kristina Toutanova. Boolq: Exploring the surprising difficulty of natural yes/no questions. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 2924–2936, 2019.

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, 2018. URL https://arxiv.org/abs/1803.05457.

Devleena Das, Rajeev Patwari, and Ashish Sirasao. Recover-lora: Data-free accuracy recovery of degraded language models via low-rank adaptation. pp. 2377–2386, 2025.

Vage Egiazarian, Andrei Panferov, Denis Kuznedelev, Elias Frantar, Artem Babenko, and Dan Alistarh. Extreme compression of large language models via additive quantization, 2024.

Elias Frantar, Saleh Ashkboos, Torsten Hoefler, and Dan Alistarh. Gptq: Accurate post-training quantization for generative pre-trained transformers, 2023. URL https://arxiv.org/abs/2210.17323.

Harsha Kokel, Michael Katz, Kavitha Srinivas, and Shirin Sohrabi. Acpbench: Reasoning about action, change, and planning. In *AAAI*. AAAI Press, 2025.

Geonho Lee, Janghwan Lee, Sukjin Hong, Minsoo Kim, Euijai Ahn, Du-Seong Chang, and Jungwook Choi. Rilq: Rank-insensitive lora-based quantization error compensation for boosting 2-bit large language model accuracy. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 18091–18100, 2025.

Jung Hyun Lee, June Yong Yang, Jungwook Choi, and Eunho Yang. LFQ: Logit-aware final-block quantization for boosting the generation quality of low-bit quantized LLMs, 2026. URL https://openreview.net/forum?id=25B8Se2kH1.

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. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=LzPWWPAdY4.

Zhen Li, Yupeng Su, SongmiaoWang, RUNMING YANG, Congkai Xie, Aofan Liu, Ming Li, Jiannong Cao, Yuan Xie, Ngai Wong, and Hongxia Yang. Quantization meets reasoning: Exploring and mitigating degradation of low-bit LLMs in mathematical reasoning, 2026. URL https://openreview.net/forum?id=So3hbnEGYV.

Hunter Lightman, Vineet Kosaraju, Yuri Burda, Harrison Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. Let's verify step by step. In B. Kim, Y. Yue, S. Chaudhuri, K. Fragkiadaki, M. Khan, and Y. Sun (eds.), *International Conference on Learning Representations*, volume 2024, pp. 39578–39601, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/file/aca97732e30bcf1303bc22ac3924fd16-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 llm compression and acceleration, 2026. URL https://arxiv.org/abs/2306.00978.

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 Aarti Singh, Maryam Fazel, Daniel Hsu, Simon Lacoste-Julien, Felix Berkenkamp, Tegan Maharaj, Kiri Wagstaff, and Jerry Zhu (eds.), *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 48371–48392. PMLR, 13–19 Jul 2025. URL https://proceedings.mlr.press/v267/patil25a.html.

Keisuke Sakaguchi, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. Winogrande: An adversarial winograd schema challenge at scale. volume 64, pp. 99–106. ACM New York, NY, USA, 2021.

Albert Tseng, Jerry Chee, Qingyao Sun, Volodymyr Kuleshov, and Christopher De Sa. QuIP#: Even better LLM quantization with hadamard incoherence and lattice codebooks. In *International Conference on Machine Learning*, 2024a. URL https://arxiv.org/abs/2402.04396.

Albert Tseng, Qingyao Sun, David Hou, and Christopher De Sa. QTIP: Quantization with trellises and incoherence processing. *arXiv preprint arXiv:2406.11235*, 2024b. URL https://arxiv.org/abs/2406.11235.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, brian ichter, Fei Xia, Ed Chi, Quoc V Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neural Information Processing Systems*, volume 35, pp. 24824–24837. Curran Associates, Inc., 2022. doi: 10.52202/068431-1800. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/9d5609613524ecf4f15af0f7b31abca4-Paper-Conference.pdf.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jing Zhou, Jingren Zhou, Junyang Lin, Kai Dang, Keqin Bao, Kexin Yang, Le Yu, Lianghao Deng, Mei Li, Mingfeng Xue, Mingze Li, Pei Zhang, Peng Wang, Qin Zhu, Rui Men, Ruize Gao, Shixuan Liu, Shuang Luo, Tianhao Li, Tianyi Tang, Wenbiao Yin, Xingzhang Ren, Xinyu Wang, Xinyu Zhang, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yinger Zhang, Yu Wan, Yuqiong Liu, Zekun Wang, Zeyu Cui, Zhenru Zhang, Zhipeng Zhou, and Zihan Qiu. Qwen3 technical report, 2025. URL https://arxiv.org/abs/2505.09388.

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*, 2023a. URL https://arxiv.org/abs/2210.03629.

Zhewei Yao, Xiaoxia Wu, Cheng Li, Stephen Youn, and Yuxiong He. Zeroquant-v2: Exploring post-training quantization in llms from comprehensive study to low rank compensation, 2023b. URL https://arxiv.org/abs/2303.08302.

Cheng Zhang, Jianyi Cheng, George Anthony Constantinides, and Yiren Zhao. LQER: Low-rank quantization error reconstruction for LLMs. In Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp (eds.), *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 58763–58779. PMLR, 21–27 Jul 2024. URL https://proceedings.mlr.press/v235/zhang24j.html.

Cheng Zhang, Jeffrey T. H. Wong, Can Xiao, George Anthony Constantinides, and Yiren Zhao. QERA: an analytical framework for quantization error reconstruction. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=LB5cKhgOTu.

A Evaluation Details and Metric Definitions

A.1 Token-Level Prediction Agreement

For a set of evaluated prediction positions $\mathcal{E}$, prediction agreement with the source model is defined as

$$\mathrm{Agree}(T, M) = \frac{1}{|\mathcal{E}|} \sum_{t \in \mathcal{E}} \mathbb{1}\left[\arg\max_v p_T(v \mid c_t) = \arg\max_v p_M(v \mid c_t)\right], \tag{5}$$

where M denotes the evaluated quantized or recovered model.

The corresponding Top-1 flip rate is

$$\mathrm{Flip}(T, M) = 1 - \mathrm{Agree}(T, M). \tag{6}$$

Answer-option agreement in the component study. The component-ablation table instead uses a decision-level statistic over the complete frozen pool of $N = 7{,}244$ multiple-choice records. For record i with answer-option set $\mathcal{O}_i$, let $s_X(i, o)$ be the raw continuation log-likelihood sum assigned by model X to option o. We define

$$\mathrm{OptionAgree}(T, M) = \frac{1}{N} \sum_{i=1}^{N} \mathbb{1}\left[\arg\max_{o \in \mathcal{O}_i} s_T(i, o) = \arg\max_{o \in \mathcal{O}_i} s_M(i, o)\right]. \tag{7}$$

The table reports $100 \times \mathrm{OptionAgree}$ in percent, with the sample standard deviation likewise in percentage points. Ties are broken by the frozen answer-option order, with the first maximum selected. Agreement is pooled over all records within each seed, after which the table reports the mean and sample standard deviation over seeds $\{42, 43, 44, 45\}$. This answer-option statistic is distinct from the vocabulary-token agreement defined above.

A.2 Jensen–Shannon Divergence

For two discrete distributions P and Q, we compute Jensen–Shannon divergence as

$$\mathrm{JSD}(P, Q) = \frac{1}{2} D_{\mathrm{KL}}(P \| M) + \frac{1}{2} D_{\mathrm{KL}}(Q \| M), \qquad M = \frac{1}{2}(P + Q). \tag{8}$$

Unless otherwise specified, JSD is reported in nats.

For the candidate-level JSD used in the controlled ablations, the candidate set is the complete set of answer options for each benchmark item. For each option, we sum the raw continuation-token log likelihoods without length normalization. We independently apply a softmax to the teacher and evaluated-model option scores and compute JSD between the resulting option distributions using natural logarithms (with a numerical probability floor of 10^{-12}).

A.3 Source Top-1 Probability Error

Let

$$v_t^* = \arg\max_v p_T(v \mid c_t) \tag{9}$$

denote the source model's Top-1 token at prediction position t. We define source Top-1 probability error as

$$E_{\mathrm{Top1}}(T, M) = \frac{1}{|\mathcal{E}|} \sum_{t \in \mathcal{E}} \left| p_M(v_t^* \mid c_t) - p_T(v_t^* \mid c_t) \right|. \tag{10}$$

This metric measures how accurately the evaluated model recovers the source model's confidence on its preferred token, independently of whether the Top-1 identity itself changes.

A.4 Teacher-Preferred Candidate Margin

For a candidate set $\mathcal{C}$ and the source-preferred candidate

$$y_T^* = \arg\max_{y \in \mathcal{C}} p_T(y \mid c), \tag{11}$$

we define the source-preferred candidate margin under model M as

$$\mathrm{Margin}_M = p_M(y_T^* \mid c) - \max_{y \in \mathcal{C} \setminus \{y_T^*\}} p_M(y \mid c). \tag{12}$$

A positive margin indicates that the source-preferred candidate remains ranked first, while a negative margin indicates a local ranking reversal.

A.5 Quantization-Induced Competitors

For the analysis in the main paper, we define the set of quantization-induced competitors at position t as

$$\mathcal{C}_t = \mathrm{TopK}_k(p_{Q_0}(\cdot \mid c_t)) \setminus \mathrm{TopK}_k(p_T(\cdot \mid c_t)). \tag{13}$$

The competitor rate is

$$R_{\mathrm{comp}} = \frac{1}{|\mathcal{E}|} \sum_{t \in \mathcal{E}} \frac{|\mathcal{C}_t|}{k}. \tag{14}$$

Using the same fixed competitor set $\mathcal{C}_t$ defined from M_Q, we measure the excess probability mass assigned by model M to these competitors as

$$E_{\mathrm{comp}}(M) = \frac{1}{|\mathcal{E}|} \sum_{t \in \mathcal{E}} \sum_{v \in \mathcal{C}_t} \max\left(p_M(v \mid c_t) - p_T(v \mid c_t), 0\right). \tag{15}$$

In the main analysis, we use $k = 32$.

A.6 Support-Budget Sensitivity

We vary the per-branch Top-k budget over $k \in \{8, 16, 32, 64\}$. Under W2, increasing k improves 0-shot Avg. from 58.59 at $k = 8$ to 62.11 at $k = 64$, with most of the improvement already realized by $k = 32$. Under W3, performance is substantially less sensitive to the support budget and remains between 70.47 and 70.58. We therefore use $k = 32$ as the default setting, which provides a compact support while capturing most of the observed gain in the more challenging W2 regime.

Table 5: **Sensitivity to the per-branch support budget k.** Results are single-seed (seed 42) Qwen3-8B QuIP# sweeps from the separate W2 and W3 controlled runs; no uncertainty estimate is available for this table.

	$k = 8$	$k = 16$	$k = 32$	$k = 64$
W2 0-shot Avg.↑	58.59	60.95	61.68	62.11
W3 0-shot Avg.↑	70.55	70.47	70.54	70.58

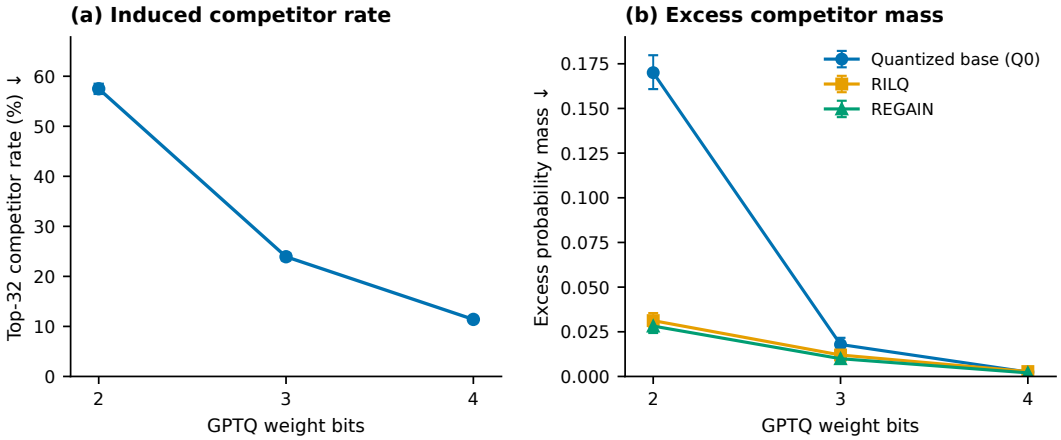

Figure 4: **Residual distortion under unchanged Top-1 predictions.** We restrict evaluation to the shared subset of prediction positions where both RILQ and REGAIN preserve the same Top-1 token as the BF16 source model. The panels report Jensen–Shannon divergence to the source distribution and absolute error in the source-model Top-1 probability under Qwen3-32B with QuIP# W2–W4. Lower is better. Even when the discrete Top-1 prediction is unchanged, measurable probability-level distortion remains, and REGAIN consistently reduces this distortion relative to RILQ.

B ADDITIONAL SAME-TOP1 ANALYSIS

The main text shows that REGAIN consistently reduces residual token-level distortion relative to RILQ. A natural question is whether this advantage comes mainly from correcting explicit Top-1 prediction flips, or whether it also persists when the discrete decision itself is unchanged. To answer this question, we perform an additional analysis on prediction positions for which the recovered model preserves the same Top-1 token as the BF16 source model.

Specifically, for each evaluated position t, let

$$y_T(t) = \arg\max_{v \in \mathcal{V}} p_T(v \mid c_t)$$

denote the source-model Top-1 token. For a recovered model $M \in \{\text{RILQ}, \text{REGAIN}\}$, we define the same-Top1 subset as

$$\mathcal{E}^M_{\text{same}} = \left\{ t : \arg\max_v p_M(v \mid c_t) = y_T(t) \right\}.$$

To enable a direct comparison between RILQ and REGAIN, we also consider the shared same-Top1 subset

$$\mathcal{E}_{\text{shared}} = \left\{ t : \arg\max_v p_{\text{RILQ}}(v \mid c_t) = y_T(t),\ \arg\max_v p_{\text{REGAIN}}(v \mid c_t) = y_T(t) \right\}.$$

All analyses below are reported on $\mathcal{E}_{\text{shared}}$ unless otherwise stated.

Figure 4 reports the residual distortion on this subset under Qwen3-32B with GPTQ W2–W4. We evaluate two complementary diagnostics: (i) Jensen–Shannon divergence to the BF16 source distribution, and (ii) absolute error in the source-model Top-1 probability. Because all positions in this subset already share the same discrete Top-1 decision as the source model, these metrics isolate purely probability-level discrepancies rather than explicit prediction flips.

The results show that measurable residual distortion remains even after conditioning on unchanged Top-1 predictions. That is, identical discrete predictions do not imply identical local token distributions. Across W2–W4, REGAIN consistently yields lower same-Top1 distortion than RILQ, indicating that its advantage is not explained solely by correcting Top-1 flips. Instead, REGAIN more faithfully restores the local probability allocation around decision-relevant competitors, including both the confidence assigned to the source-preferred token and the relative mass assigned to nearby alternatives.

The effect is strongest under W2, where aggressive quantization causes the largest distortion. These observations support the main claim of the paper: recovering the discrete Top-1 token alone is insufficient to recover the source model's local output behavior, and quantization-aware compact supervision improves recovery precisely at this probability-distribution level.

C QUALITATIVE EXAMPLES

The motivating examples in Figure 1 are selected from the same benchmark distributions used in the quantitative evaluation. For each example, we report the dataset split and example identifier together with the outputs of the BF16 source model, the initial quantized model M_Q, and REGAIN. The examples are deliberately mechanism-selected rather than sampled at random. The ARC-Challenge case was chosen from the GPTQ-W3 static-score records because M_Q retained the correct top-1 answer despite a strong shift in the four-candidate distribution, while REGAIN moved the distribution closer to BF16. For ACPBench, we first retained 86 examples whose decoded outputs were correct for BF16, incorrect for M_Q, and correct for REGAIN. We then required the action-likelihood top-1 to exhibit the same pattern (five examples), preferred agreement between decoded labels and likelihood top-1, and ranked the remaining cases by the reduction in JSD to BF16, using the M_Q–BF16 JSD as the final tie-breaker. These examples are therefore diagnostic rather than representative samples.

C.1 MULTIPLE-CHOICE EXAMPLE

The multiple-choice example in Figure 1(a) comes from ARC-Challenge, validation split, example ID `Mercury_7179358`.

Table 6: **Metadata for the multiple-choice motivating example.**

Field	Value
Dataset	ARC-Challenge (`arc_challenge`)
Split	validation
Example ID	`Mercury_7179358`
Question	A climax fire ecosystem requires periodic forest fires to maintain stability. Which of these would be the most likely result of preventing natural fires from occurring in this ecosystem?
Candidate A	Pine species would reproduce more rapidly.
Candidate B	Broadleaf species would replace pine species.
Candidate C	Burning areas would be more easily contained.
Candidate D	Trees would spread into previously unforested areas.
Reference answer	B: Broadleaf species would replace pine species.
BF16 prediction	B (continuation-likelihood top-1; correct)
M_Q prediction	B (continuation-likelihood top-1; correct)
REGAIN prediction	B (continuation-likelihood top-1; correct)

The candidate scores used in Figure 1(a) are reported in Table 7.

Table 7: **Candidate probabilities for the multiple-choice motivating example.**

Model	A	B	C	D
BF16 Source	0.000433	0.998412	2.33×10^{-7}	0.001155
M_Q	0.428102	0.564104	0.000663	0.007131
REGAIN	0.082280	0.916540	8.37×10^{-6}	0.001171

C.2 STRUCTURED-DECISION EXAMPLE

The structured-decision example in Figure 1(b) comes from ACPBench, test split, example ID `−3885269214599382873` (sample `69`).

Table 8: **Metadata for the ACPBench motivating example.**

Field	Value
Dataset	ACPBench
Split	test
Example ID	`-3885269214599382873` (sample `69`)
Task / environment	`acp_local_just_mcq`; action-justification MCQ; Rovers
Decision context	Given an 11-action Rovers plan whose goals are to communicate rock data, soil data, and a colour image, identify which candidate action can be removed while the remaining sequence is still a valid plan.
Source-preferred action	C: move rover1 from waypoint2 to waypoint0
M_Q selected action	D: take a colour image of objective1 with camera2 on rover1 from waypoint2
REGAIN selected action	C: move rover1 from waypoint2 to waypoint0
BF16 evaluator outcome	Correct (decoded "C."; exact match)
M_Q evaluator outcome	Incorrect (decoded "D."; not an exact match)
REGAIN evaluator outcome	Correct (decoded "C"; exact match)

Table 9: **Local action probabilities for the ACPBench motivating example.**

Candidate action	BF16 Source	M_Q	REGAIN
A: sample rock at waypoint0 with rover1 and store it in store1	0.016899	0.026576	0.031577
B: communicate rover0's soil data from waypoint0 to the lander at waypoint1	0.029627	0.003885	0.052362
C: move rover1 from waypoint2 to waypoint0	0.728816	0.340240	0.638761
D: take a colour image of objective1 with camera2 on rover1 from waypoint2	0.224658	0.629299	0.277300

At the decision-relevant position shown in Figure 1(b), the local candidate probabilities are:

All probabilities in the two qualitative tables are obtained by applying a softmax over the four raw continuation log-likelihoods (answer continuations for ARC and action-text continuations for ACPBench). They are diagnostic candidate-preference distributions, not official ACPBench accuracy probabilities.

The example illustrates the distinction emphasized in the main paper: quantization can promote a candidate that was not locally preferred by the source model, and a resulting local ranking change can alter the structured decision. REGAIN explicitly includes such quantization-induced competitors in its compact support, allowing the recovery objective to directly constrain their probability mass.