DOES QUANTIZATION PRESERVE HOW LANGUAGE MODELS ANSWER? CAUSAL-EQUIVARIANT QUANTIZATION FOR SYMBOL-BINDING CIRCUITS

Anonymous authors
Paper under double-blind review

ABSTRACT

Post-training quantization (PTQ) is typically optimized to preserve weights and activations while maintaining perplexity, or task accuracy. However, these objectives do not reveal whether quantization preserves the causal computation underlying model decisions. Recent studies have localized answer selection and symbol binding in formatted multiple-choice question answering (MCQA) to a sparse set of attention components, making it possible to directly examine their preservation under quantization. To preserve the causal mechanisms underlying formatted MCQA after quantization, we propose Causal-Equivariant Quantization (CEQ), a structured mixed-precision framework consisting of three components: counterfactual causal probing, causal-regret-based precision allocation, and causal-equivariant calibration. First, counterfactual causal probing identifies model components whose causal behavior is most affected by low-bit quantization across different answer formats. CEQ measures these changes using causal-effect distortion (CED) and format-equivariance error (FEE). Second, causal-regret-based precision allocation uses this information to assign higher precision to causally important and quantization-sensitive components under a global bit budget. Third, causal-equivariant calibration (CEQ-CAL) further reduces the discrepancy between the quantized and full-precision models while maintaining consistent behavior across answer formats by preserving output distributions, answer margins, causal effects, and format equivariance. We evaluated CEQ on a novel controlled symbol-binding dataset constructed specifically for this work and on five public MCQA datasets. On the controlled symbol-binding benchmark, CEQ achieves 83.40 ± 14.21% worst-format accuracy at 2.72 effective matrix bits versus 68.00 ± 20.64% for uniform 2-bit RTN. Across the five public MCQA datasets, CEQ consistently improves the worst-format accuracy over the matched TaCQ baseline, achieving an average gain of 9.86% points. The effectiveness of CEQ further generalizes across model scales and architecture families, achieving 76.55% MMLU accuracy on Qwen2.5-32B-Instruct at 2.12 bits and a 67.95% zero-shot average on Llama-3-8B at 4.08 bits. These results show that preserving prediction quality under quantization does not necessarily preserve the underlying causal computation, and motivate mechanism preservation as a distinct objective for low-bit language-model compression.

1 INTRODUCTION

Large language models are increasingly deployed after aggressive low-bit compression. Modern PTQ techniques can preserve downstream quality by minimizing local reconstruction error, protecting salient weights, rotating activation outliers, or assigning mixed precision to sensitive subspaces (Frantar et al., 2023; Lin et al., 2025; Xiao et al., 2023; Ashkboos et al., 2024; Liu et al., 2025; Huang et al., 2025). Yet output preservation and mechanism preservation are different objectives. A quantized model can remain correct on familiar prompts while relying on a different internal computation, potentially weakening robustness when the task is expressed through an equivalent but less frequent format.

Formatted MCQA provides an unusually clean setting in which to test this possibility. A model must first identify the semantic answer and then map that answer to a prompt-specific symbol. Wiegreffe

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.

et al. (2025) showed that successful models implement this behavior through a sparse sequence of attention-mediated operations. Middle layers are causally important for selecting the relevant answer, later layers promote a vocabulary token, and unusual answer alphabets can trigger a delayed symbol-alignment stage. Their results also show that only a small number of heads per layer contribute strongly to answer-symbol production. This suggests a concrete compression question: *if a capability is sparse and causally localized, should precision be allocated according to how quantization changes that causal role?*

We operationalize this question by measuring how quantization alters a component's intervention effect across semantically equivalent prompt formats. Existing mixed-precision methods do not directly optimize quantization-induced changes in a component's causal contribution. TaCQ (Xiao et al., 2025), use task-conditioned gradients together with predicted quantization perturbations to identify weights whose quantization is likely to adversely affect task performance. CoQuant (Ding et al., 2026) protects a joint weight–activation subspace. These methods establish that task-aware or subspace-aware allocation is useful, but they do not explicitly preserve intervention effects under known semantic equivalences. Our central hypothesis is therefore narrower: when a task admits controlled counterfactual transformations, the full-precision model itself provides a causal reference signal for deciding which compressed components must retain precision.

Our core idea is to make mechanism preservation an explicit quantization objective by measuring how quantization changes component-level causal effects across semantically equivalent prompts and using these measurements to guide precision allocation and calibration. We introduce Causal-Equivariant Quantization (CEQ), which uses controlled changes in answer positions and symbol alphabets to compare component-level intervention effects between full-precision and low-bit models. These comparisons yield a causal quantization regret score that identifies causally important components whose effects are sensitive to quantization, guiding their protection at higher precision under a global bit budget. After allocation, CEQ-CAL adjusts quantizer parameters to preserve the full-precision model's output distributions, answer margins, and intervention effects while encouraging consistent behavior across equivalent answer formats.

We make five contributions. (1) We propose CEQ, a mechanism-preserving mixed-precision quantization framework for low-bit language models. (2) We introduce counterfactual causal probing across semantically equivalent answer formats, together with CED and FEE to quantify causal and equivariance distortion. (3) We develop causal-regret-based precision allocation to protect quantization-sensitive attention groups under a global bit budget. (4) We introduce CEQ-CAL, which calibrates quantizer parameters to preserve output distributions, answer margins, causal effects, and format equivariance. (5) We develop a controlled symbol-binding benchmark and conduct extensive evaluations across matched-budget baselines, architecture variants, ablations, exhaustive allocation, and pretrained Qwen2.5 and Llama3 models.

2 BACKGROUND AND RELATED WORK

2.1 MCQA MECHANISMS.

Formatted MCQA requires a model to identify the semantically correct answer and bind it to a prompt-specific symbol. Behavioral studies demonstrate sensitivity to answer ordering and label representation (Pezeshkpour and Hruschka, 2024; Zheng et al., 2024; Sun et al., 2024). Complementing these studies, mechanistic analyses identify a sparse set of answer-letter heads and associated components (Lieberum et al., 2023). Most directly relevant to our work, Wiegreffe et al. (2025) combine activation patching with vocabulary projection across OLMo, Llama, and Qwen, using Colors to disentangle semantic knowledge from symbol binding. We investigate whether this localized computation is preserved under low-bit quantization perturbations.

2.2 QUANTIZATION AND TASK-AWARE ALLOCATION.

GPTQ employs approximate second-order reconstruction (Frantar et al., 2023), while AWQ uses activation salience to guide quantization (Lin et al., 2025). SqueezeLLM and SpQR retain sensitive or outlier weights at higher precision (Kim et al., 2024; Dettmers et al., 2023), whereas SliM-LLM allocates mixed precision according to salience (Huang et al., 2025). Complementing selective precision, QuaRot and SpinQuant apply rotations (Ashkboos et al., 2024; Liu et al., 2025). Calibration-oriented

methods, including OmniQuant and FlatQuant, optimize equivalent transformations or clipping and scaling parameters (Shao et al., 2024; Sun et al., 2025). Beyond numerical sensitivity, TaCQ (Xiao et al., 2025) identifies task-relevant weights using task gradients and predicted quantization perturbations. CoQuant instead protects a joint weight–activation subspace (Ding et al., 2026).

These approaches do not directly optimize the preservation of causal intervention effects under known semantic equivalences. CEQ uses controlled transformations to specify the required invariance and causal interventions to identify internal components whose functional effects must be preserved. Table 1 summarizes this methodological distinction.

Table 1: Conceptual positioning of the quantization baselines. "Intervention" means that the precision signal directly depends on intervening on an internal component; "equivariance" means semantics-preserving prompt transformations enter the objective.

Method	Primary signal	Mixed precision	Intervention	Equivariance
GPTQ	second-order reconstruction	optional	✗	✗
AWQ	activation salience	optional	✗	✗
SqueezeLLM	sensitivity + sparse outliers	✓	✗	✗
SpQR	second-order + sparse outliers	✓	✗	✗
QuaRot	rotation / outlier suppression	✗	✗	✗
SpinQuant	learned rotations	✗	✗	✗
SliM-LLM	salience-driven bit allocation	✓	✗	✗
TaCQ	task-localized sensitivity	✓	✗	✗
CoQuant	joint weight–activation subspace	✓	✗	✗
CEQ	counterfactual causal regret	✓	✓	✓

3 Problem Formulation

Let f be a full-precision autoregressive LM and $Q_{\mathbf{b}}(f)$ a quantized model whose structured groups use bit widths $\mathbf{b}$. For an MCQA instance x, let $z_f(x) \in \mathbb{R}^K$ denote logits restricted to the K answer symbols. A transformation τ changes answer labels and/or answer positions while preserving the semantic answer. Since the correspondence between the canonical and transformed answer labels is known, P_τ denotes the permutation that maps canonical answer coordinates to their corresponding coordinates in the transformed output space.

Format-equivariance error. A robust selector should transform consistently with the prompt. We define:

$$\mathrm{FEE}(Q) = \mathbb{E}_{x,\tau}\left[\left\|P_\tau^{-1}\pi(Q,\tau(x)) - \pi(Q,x)\right\|_2^2\right]. \tag{1}$$

where π is the softmax distribution over the answer candidates. FEE is intentionally different from ordinary accuracy: a model can predict the correct answer on both prompts while assigning different relative evidence to equivalent positions.

Causal-effect distortion. For component c, let I_c be an intervention such as zeroing or replacement of the activation. For a scalar answer margin $m_f(x)$, define the full-precision intervention effect

$$\Delta_c^f(x) = m_f(x) - m_{I_c(f)}(x). \tag{2}$$

The quantized analogue is $\Delta_c^Q(x)$. We define CED over the component set $\mathcal{C}$, evaluation set $\mathcal{D}$, and transformation family $\mathcal{T}$ as:

$$\mathrm{CED}(Q,f) = \left[\frac{1}{|\mathcal{C}||\mathcal{D}||\mathcal{T}|}\sum_{c\in\mathcal{C}}\sum_{x\in\mathcal{D}}\sum_{\tau\in\mathcal{T}}\left(\Delta_c^Q(\tau(x)) - \Delta_c^f(\tau(x))\right)^2\right]^{1/2}. \tag{3}$$

CED quantifies the root-mean-square deviation between intervention effects in the quantized and full-precision models. It therefore assesses whether quantization preserves the causal contribution of each evaluated internal component, beyond preserving the final predicted label.

4 Causal-Equivariant Quantization

As illustrated in Fig. 1, CEQ comprises three stages. First, counterfactual causal probing compares intervention effects in the full-precision and low-bit models across semantically equivalent prompt

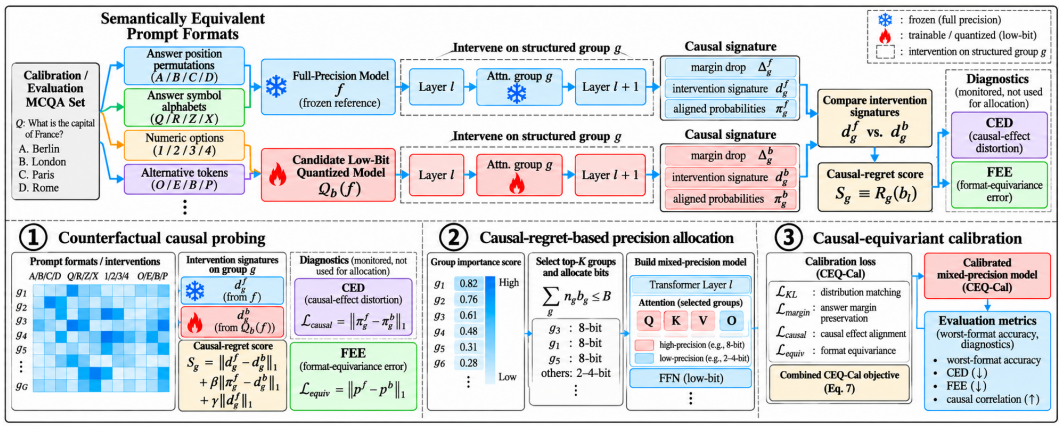

Figure 1: Overview of the proposed CEQ framework. **Top:** We construct semantically equivalent prompt formats and intervene on the same structured group g in the full-precision model f and the candidate low-bit model $Q_{b_l}(f)$. Their intervention signatures, d_g^f and $d_g^{b_l}$, are compared to compute the causal-regret protection score $S_g \equiv R_g(b_l)$ in Eq. 4. **Bottom:** CEQ proceeds through three stages: (1) counterfactual causal probing, (2) causal-regret-based precision allocation under a global bit budget, and (3) causal-equivariant calibration (CEQ-Cal).

formats to characterize quantization-induced changes in causal behavior. Second, causal-regret-based precision allocation uses these changes to identify quantization-sensitive attention groups and assign them higher precision under a global bit budget. Third, causal-equivariant calibration (CEQ-Cal) adjusts quantizer parameters to preserve output distributions, answer margins, causal effects, and consistency across equivalent prompt formats.

4.1 Counterfactual Causal Probing and Quantization Regret

For each structured group g, we first evaluate its vulnerability under the candidate low-bit quantization b_l. For a model instance M, let $d_g(M)$ denote the intervention signature of structured group g across answer alphabets and answer positions. We write $d_g^f \equiv d_g(f)$ for the full-precision model and $d_g^{b_l} \equiv d_g(Q_{b_l}(f))$ for the candidate low-bit model. We define the causal-regret protection score as:

$$S_g \equiv R_g(b_l) = \underbrace{\|d_g^f - d_g^{b_l}\|_1}_{\text{causal distortion}} + \beta \underbrace{\|\bar{d}_g^f - \bar{d}_g^{b_l}\|_1}_{\text{role/equivariance distortion}} + \gamma \underbrace{\|d_g^f\|_1}_{\text{causal strength}} . \tag{4}$$

we use $\beta = 0.5$ and $\gamma = 0.10$ for CEQ causal-regret scoring. These values are fixed across the reported CEQ experiments and are not selected using held-out evaluation accuracy. Appendix K.1 varies β only as a sensitivity analysis while keeping $\gamma = 0.10$ fixed; the default $\beta = 0.5$ is retained for the main reported results. Here, $\bar{d}_g^f$ and $\bar{d}_g^{b_l}$ denote the centered full-precision and low-bit intervention signatures, respectively, obtained by subtracting the mean intervention effect across the considered formats. Centering isolates format-dependent variation from the component's overall causal-effect magnitude.

The first term identifies components whose causal function is strongly distorted by low-bit quantization; the second emphasizes differential behavior across semantically equivalent formats; and the third prioritizes components that have a large causal contribution in the full-precision model. Importantly, S_g is a protection score rather than a per-bit objective: a high score indicates that leaving group g at low precision is undesirable.

4.2 Causal-Regret-Based Precision Allocation

Given these low-bit protection scores, CEQ formulates mixed-precision allocation as a budgeted protection problem. Let $z_g \in \{0, 1\}$ indicate whether group g is protected at the higher precision b_h ($z_g = 1$) or retained at the low precision b_l ($z_g = 0$). We minimize the total protection score of groups that remain exposed to low-bit quantization:

$$\min_{\{z_g\}} \sum_g (1 - z_g) S_g \quad \text{s.t.} \quad \sum_g n_g \left[b_l + z_g(b_h - b_l)\right] \le B, \qquad z_g \in \{0, 1\}. \tag{5}$$

Algorithm 1 Two-level CEQ allocation

Require: FP model f, calibration set $\mathcal{D}$, transforms $\mathcal{T}$, low/high bits (b_l, b_h), head budget K

1: Construct equivalent prompts by varying answer positions and symbol alphabets.
2: Measure full-precision causal signatures $\mathbf{d}_g^f$.
3: Quantize all candidate groups to b_l; measure $\mathbf{d}_g^q$.
4: Compute the causal-regret protection score $S_g = R_g(b_l)$ using Eq. 4.
5: Solve Eq. 5; for equal-size groups and a K-group budget, protect the K groups with the largest S_g at b_h and retain b_l elsewhere.
6: **return** mixed-precision quantized model and allocation metadata.

This objective uses S_g as a low-bit vulnerability score; it does not assume that a protected group has exactly zero quantization error. Rather, protection removes that group from the low-bit backbone. When all candidate groups have equal size $n_g = n$, the global budget B permits

$$K = \left\lfloor \frac{B - |\mathcal{G}| n b_l}{n(b_h - b_l)} \right\rfloor$$

groups to be protected at b_h, where $|\mathcal{G}| n b_l$ is the cost of keeping all candidate groups with low precision b_l. Equation 5 then reduces to selecting the K groups with the largest S_g, as used in Algorithm 1 and in our controlled-dataset experiments. For heterogeneous group sizes, Eq. 5 becomes a standard budgeted binary selection problem.

4.3 Causal-equivariant calibration

After allocation, CEQ-Cal optimizes only quantizer parameters (e.g., clipping/scales), not base-model weights. For each loss component $j \in \mathcal{J} = \{\mathrm{KL}, \mathrm{margin}, \mathrm{causal}, \text{equiv-ref}, \mathrm{FEE}\}$, we normalize its value over the clipping-factor grid $\mathcal{C} = \{0.65, 0.70, \ldots, 1.05\}$ as:

$$\widetilde{\mathcal{L}}_j(\alpha) = \frac{\mathcal{L}_j(\alpha)}{\max(\mathrm{median}_{\alpha' \in \mathcal{C}}\, \mathcal{L}_j(\alpha'), 10^{-12})}. \tag{6}$$

The CEQ-Cal objective used in our experiments is:

$$\begin{aligned} \mathcal{L}_{\text{CEQ-CAL}}(\alpha) = \widetilde{\mathcal{L}}_{\mathrm{KL}}(\alpha) + 0.25\, \widetilde{\mathcal{L}}_{\mathrm{margin}}(\alpha) + \widetilde{\mathcal{L}}_{\mathrm{causal}}(\alpha) \\ + 0.50\, \widetilde{\mathcal{L}}_{\mathrm{equiv-ref}}(\alpha) + 0.125\, \widetilde{\mathcal{L}}_{\mathrm{FEE}}(\alpha). \end{aligned} \tag{7}$$

The final clipping factor is selected on the calibration set as $\hat{\alpha} = \arg\min_{\alpha \in \mathcal{C}} \mathcal{L}_{\text{CEQ-CAL}}(\alpha)$. Here, $\widetilde{\mathcal{L}}_{\mathrm{KL}}$ matches the FP answer distribution, $\widetilde{\mathcal{L}}_{\mathrm{margin}}$ matches the correct-vs. the strongest-wrong margin, and $\widetilde{\mathcal{L}}_{\mathrm{causal}}$ matches the intervention signature. $\widetilde{\mathcal{L}}_{\mathrm{equiv-ref}}$ measures the mean squared difference between the FP and quantized candidate probability distributions, while $\widetilde{\mathcal{L}}_{\mathrm{FEE}}$ measures the inconsistency between formats in the calibration set. The coefficients are fixed for all five seeds; only the global clipping factor is selected on the calibration set.

5 Experimental Study

5.1 Datasets and evaluation suite

We evaluate CEQ on a controlled symbol-binding dataset and five public multiple-choice question answering (MCQA) datasets: Colors, MMLU, HellaSwag, ARC-Challenge, and CommonsenseQA. Together, these datasets cover controlled symbol binding, context-dependent answer mapping, multidomain knowledge, common sense reasoning, and science question answering. This evaluation suite allows us to test CEQ at two complementary levels: controlled mechanistic analysis, where causal interventions can be performed exhaustively, and pretrained LLM evaluation, where CEQ is assessed on realistic reasoning benchmarks under standardized counterfactual answer-format transformations.

Public MCQA benchmarks. We further evaluate CEQ on five public benchmarks using pretrained language models. Colors is a controlled copying-from-context task that isolates the mapping between semantic content and prompt-specific response symbols (Wiegreffe et al., 2025; Norlund et al., 2021). MMLU contains 57 academic and professional subject areas and evaluates broad knowledge and problem solving (Hendrycks et al., 2021). HellaSwag is a four-way adversarial commonsense-completion benchmark designed to distinguish robust commonsense reasoning from shallow continuation heuristics (Zellers et al., 2019). ARC-Challenge contains difficult grade-school science questions selected to challenge retrieval- and co-occurrence-based methods (Clark et al., 2018). CommonsenseQA is a five-choice benchmark derived from ConceptNet relations and evaluates commonsense reasoning beyond surface-level lexical associations (Talmor et al., 2019).

For each four-choice benchmark, we construct counterfactual variants by crossing four answer positions with four answer-symbol alphabets, while preserving the underlying semantic answer. This produces 16 equivalent prompt forms for each semantic example. For CommonsenseQA, which contains five candidate answers, the same protocol is extended to $K = 5$ using five answer positions and five-symbol alphabets. CEQ is evaluated on all transformed variants using the same prompt templates, calibration protocol, and evaluation splits as the corresponding baselines. We report conventional task accuracy together with worst-format accuracy, Format-Equivariance Error (FEE), Causal-Effect Distortion (CED), and causal-effect preservation metrics. This design directly tests whether CEQ preserves not only task performance, but also robust answer-selection behavior under semantically equivalent changes in answer representation. A compact dataset summary, including the exact calibration and evaluation sizes, split definitions, and number of counterfactual prompt variants, is provided in Appendix 2 and Table 2.

5.2 Controlled Benchmark Setup

We evaluate CEQ on three causal-Transformer variants: Base-3L, Deep-4L, and Wide-3L, using five independently trained seeds per architecture. Base-3L contains three layers, four attention heads per layer, $d_{\text{model}} = 32$, and FFN dimension 64; Deep-4L adds one Transformer block, while Wide-3L uses three layers, five heads, $d_{\text{model}} = 40$, and FFN dimension 80. Models are trained procedurally for 360 steps with batch size 128 using AdamW.

Causal discovery uses 60 semantic specifications together with four position-controlled subsets of 15 examples. Evaluation uses 500 held-out semantic specifications per seed under the four answer alphabets `A/B/C/D`, `Q/R/Z/X`, `1/2/3/4`, and `O/E/B/P`, yielding 2,000 paired formatted prompts for each model–method configuration.

All linear transformations are quantized using symmetric round-to-nearest (RTN) quantization, while embeddings, normalization parameters, and biases remain in full precision. We evaluate uniform 4-, 3-, and 2-bit settings using the symmetric RTN operator defined in Eq. 1; under this convention, the 2-bit setting is a three-level (ternary) symmetric quantizer. We also evaluate mixed-precision variants in which selected attention-head Q/K/V/O slices are retained at 8 bits. Our principal stress setting uses this ternary 2-bit backbone with three protected heads, corresponding to 2.72 effective matrix bits.

To isolate the effect of CEQ's allocation criterion, we compare against activation-, magnitude-, gradient-, Fisher-, and random-based head selection under the same quantizer, calibration set, precision budget, and evaluation prompts. Thus, the only experimental variable is the rule used to select higher-precision components.

For each quantized model, we report average accuracy, worst-format accuracy, Format-Equivariance Error (FEE), Causal-Effect Distortion (CED), causal-effect correlation, and effective matrix bits. All comparisons are paired across the same semantic specifications and are reported as mean $\pm$ standard deviation over five independent seeds.

5.3 Main results

Uniform 3-bit RTN is already nearly saturated on this small task, reaching $99.48 \pm 0.69\%$ worst-format accuracy. This ceiling is useful as a sanity check but makes selection criteria difficult to distinguish. Figure 2 illustrates the relationship between effective matrix precision and worst-format accuracy, highlighting the greater separation among methods in the aggressive low-bit regime. Uniform RTN-2

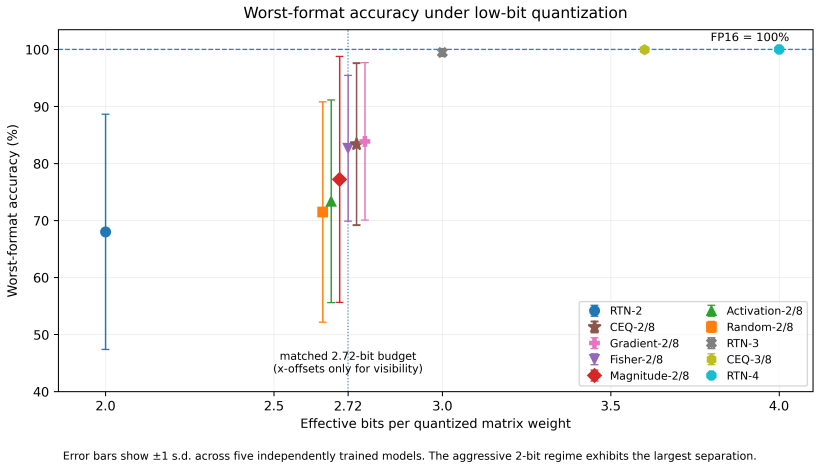

Figure 2: Worst-format accuracy under low-bit quantization. Results are averaged over five independently trained models, with error bars denoting ± 1 standard deviation.

Table 2: Equal-budget stress test at 2.72 effective matrix bits (mean $\pm$ s.d., five independently trained models)

Selection	Bits	Worst acc. (%) ↑	FEE ↓	CED ↓	Causal corr. ↑
CEQ-2/8	2.72	83.40 ± 14.21	0.0527 ± 0.0369	0.465 ± 0.075	0.833 ± 0.074
Gradient-2/8	2.72	83.88 ± 13.80	0.0466 ± 0.0270	0.465 ± 0.108	0.827 ± 0.093
Fisher-2/8	2.72	82.68 ± 12.78	0.0459 ± 0.0268	0.482 ± 0.093	0.821 ± 0.092
Magnitude-2/8	2.72	77.20 ± 21.56	0.0629 ± 0.0315	0.489 ± 0.093	0.785 ± 0.073
Activation-2/8	2.72	73.36 ± 17.77	0.0778 ± 0.0335	0.530 ± 0.096	0.759 ± 0.085
Random-2/8	2.72	71.48 ± 19.32	0.0749 ± 0.0358	0.610 ± 0.061	0.683 ± 0.067

falls to $68.00 \pm 20.64\%$ worst-format accuracy with CED 0.605 ± 0.079. Protecting three heads selected by CEQ increases the effective matrix precision to 2.72 bits and raises worst-format accuracy to $83.40 \pm 14.21\%$, while reducing CED to 0.465 ± 0.075.

Table 2 compares the allocation criteria at a matched budget of 2.72 effective matrix bits. CEQ improves mean worst-format accuracy over activation ($+10.04$ points), magnitude ($+6.20$), and random ($+11.92$) selection. The paired five-seed differences are positive in all five seeds against activation and random selection. Gradient selection is slightly higher on mean worst-format accuracy (83.88% vs. 83.40%), while CEQ has marginally higher causal correlation; the difference is not statistically resolvable at five seeds. This is an important boundary on the claim: the controlled experiment supports causal-regret allocation over weak salience heuristics, but does not establish universal superiority over strong task-gradient/Fisher criteria.

5.4 Pretrained-LLM Experimental Results

We next evaluate CEQ on pretrained Qwen2.5 and Llama3 models to determine whether the causal-equivariant allocation principle identified in the controlled study transfers to realistic language-model settings. We compare CEQ with representative reconstruction-, salience-, task-, and rotation-based quantization methods under comparable low-bit operating regimes. The nominal bit widths in the pretrained-model comparisons denote the effective weight-precision budgets of the corresponding methods and should not be interpreted as implying an identical quantization alphabet or quantization operator across methods. In particular, the three-level $b = 2$ RTN operator used in the controlled study is specific to that matched-quantizer experiment. Extended pretrained-model settings and supplementary result tables are provided in Appendix 13.

Qwen2.5 results. Table 3 reports MMLU performance on Qwen2.5-7B-Instruct. The aggressive 2-bit regime exposes substantial differences between quantization strategies. CEQ achieves 63.40% at 2.12 effective bits, improving over TaCQ by 7.27 percentage points while using a closely matched precision budget.

At approximately 3 bits, the gap between methods narrows as quantized models approach the FP16 performance ceiling. CEQ reaches 72.55% at 3.12 effective bits, reducing the gap to the 73.90%

Table 3: Low-bit quantization results on Qwen2.5-7B-Instruct. Avg. bits denotes effective average weight precision.

	∼2-bit		∼3-bit	
Method	**Bits**	**MMLU (%) ↑**	**Bits**	**MMLU (%) ↑**
FP16	16.00	73.90	16.00	73.90
GPTQ	2.00	38.17	3.00	69.14
SliM-LLM	2.125	47.89	3.125	70.42
SqueezeLLM	2.15	36.55	3.145	69.25
SpQR	2.15	25.33	3.14	67.87
TaCQ	2.10	56.13	3.12	71.64
CEQ	**2.12**	**63.40**	**3.12**	**72.55**

Table 4: Public MCQA results on Qwen2.5-7B-Instruct. Acc. denotes conventional task accuracy and Worst denotes the minimum accuracy across semantics-preserving answer-format variants. TaCQ and CEQ operate at approximately 2.10 and 2.12 effective bits, respectively.

	FP16		TaCQ		CEQ	
Dataset	Acc.	Worst	Acc.	Worst	Acc.	Worst
Colors	96.19	91.43	82.86	70.48	**89.52**	**81.90**
MMLU	73.90	68.15	56.13	45.72	**63.40**	**55.26**
HellaSwag	80.12	75.46	64.28	53.91	**71.36**	**63.48**
ARC-Challenge	61.94	56.88	45.72	36.40	**52.81**	**45.73**
CommonsenseQA	72.89	67.41	55.61	45.37	**62.95**	**54.82**

FP16 model to only 1.35 percentage points. The substantially larger gain in the 2-bit regime indicates that causal-equivariant allocation is most beneficial when the precision budget is highly constrained.

The trend remains visible at larger scale. CEQ achieves 76.55% MMLU at 2.12 effective bits, improving over TaCQ by 4.04 percentage points at a closely matched precision budget (see Appendix 13.3 and Table 25). This further illustrates the sensitivity of large pretrained models to component selection under aggressive quantization.

Public MCQA benchmark results. To determine whether the MMLU result generalizes beyond a single benchmark, we evaluate Qwen2.5-7B-Instruct on the complete public MCQA dataset using the same counterfactual answer-format protocol described in Section 5.1. Table 4 reports the conventional and worst-format accuracy for FP16, TaCQ, and CEQ on Colors, MMLU, HellaSwag, ARC-Challenge, and CommonsenseQA. Detailed results on FEE, CED, and causal-effect preservation are reported in the Appendix 13.5.

Llama3 results. We evaluate CEQ on the pretrained Meta Llama 3 8B model (`meta-llama/Meta-Llama-3-8B`). Table 5 reports the corresponding 4-bit quantization results. CEQ achieves 67.95% at 4.08 effective bits, exceeding the strongest competing quantized baseline, QuaRot, by 0.68 percentage points and retaining 99.8% of the FP16 downstream score. The improvement in downstream accuracy therefore does not come at the cost of degraded language-modeling quality.

CEQ's causal probing is confined to offline model preparation and introduces no intervention operations during deployment. Detailed selection time, calibration time, peak GPU memory, and decoding-throughput measurements for Qwen2.5-7B, Qwen2.5-32B, and Llama-3-8B are reported in Appendix M.6.

5.5 Architecture Generalization

We repeat the five-seed 2/8-bit comparison on Base-3L, a deeper 4-layer/4-head model (Deep-4L), and a wider 3-layer/5-head model with $d_{\text{model}} = 40$ (Wide-3L). Relative to RTN-2, CEQ improves worst-format accuracy by 15.4, 18.7, and 17.4 percentage points, respectively. In contrast, CEQ remains statistically indistinguishable from gradient selection ($+0.24$ points, $p = 0.403$) and Fisher selection ($+0.93$ points, $p = 0.216$), identifying these task-sensitivity criteria as the strongest competing explanations. Full architecture-resolved results are provided in Appendix 7.

Table 5: 4-bit quantization results on Llama-3-8B.

Method	Precision	Zero-shot Avg. (%) ↑	WikiText-2 PPL ↓
FP16	16-bit	68.09	6.14
GPTQ	4-bit	61.03	7.43
AWQ	4-bit	67.03	7.36
QuaRot	4-bit	67.27	6.53
SpinQuant	4-bit	66.54	6.49
CEQ	**4.08-bit**	**67.95**	**6.37**

5.6 Do FEE and CED predict robustness?

Pooling 105 non-FP16 quantized model–method instances across the three architecture families, FEE is strongly negatively associated with worst-format accuracy (Pearson $r = -0.826$, $p = 2.35 \times 10^{-27}$; Spearman $\rho = -0.869$, $p = 2.54 \times 10^{-33}$), as is CED (Pearson $r = -0.661$, $p = 1.64 \times 10^{-14}$; Spearman $\rho = -0.705$, $p = 4.38 \times 10^{-17}$). Causal-effect correlation is positively associated with worst-format accuracy (Pearson $r = 0.617$, $p = 2.50 \times 10^{-12}$; Spearman $\rho = 0.673$, $p = 3.86 \times 10^{-15}$). These are observational correlations, not causal estimates; architecture-resolved plots and statistics are reported in Appendix 9.

5.7 Ablation study

The main paper reports three ablations that directly test the CEQ design. The number of protected heads, the weight β on the cross-format regret term, and the calibration objective. Relative to the uncalibrated CEQ allocation, full CEQ-Cal raises worst-format accuracy from $83.40 \pm 14.21\%$ to $95.16 \pm 4.35\%$, reduces FEE from 0.0527 to 0.0156, lowers CED from 0.465 to 0.365, and increases causal-effect correlation from 0.833 to 0.895. KL-only calibration achieves slightly higher mean worst-format accuracy (95.44%) but lower causal-effect correlation (0.884), revealing the tradeoff between predictive robustness and causal-effect preservation. This supports the multi-objective design of CEQ-Cal, which jointly optimizes behavioral accuracy and preservation of the underlying causal mechanism. A compact ablation summary and interpretation are provided in Appendix 3. Appendix 6 provides detailed analyses of calibration-set size, selector/binder/promoter role scores, paired bootstrap intervals, circuit non-additivity, architecture-wise statistics, format transfer, layerwise regret, exhaustive enumeration of all 220 Base-3L three-head allocations, and negative interaction-aware selection results.

6 Discussion

Our experiments consistently support the central hypothesis of CEQ that effective low-bit quantization depends not only on preserving numerical reconstruction quality, but also on preserving the internal computations that determine model behavior. In the controlled symbol-binding study, CEQ improves robustness over uniform RTN and weaker salience-based allocation strategies under the same precision budget, while substantially reducing FEE and CED. The strong correlation between these diagnostic metrics and worst-format accuracy further indicates that the proposed causal and equivariance objectives capture behaviorally meaningful quantization damage rather than merely parameter-level perturbation. The corresponding quantitative comparisons are reported in Table 4 and Appendix 13.5.

7 Conclusion

Does quantization preserve how language models answer? Our results show that the answer is *"not necessarily"*. A quantized model can retain strong task accuracy while still altering the causal computation and format-equivariant behavior that produced the full-precision prediction. This distinction motivates CEQ, which allocates precision according to quantization-induced distortion of counterfactual causal effects rather than numerical reconstruction alone. CED and FEE make this loss of mechanism directly measurable, while CEQ-Cal extends the same principle to quantizer calibration. Taken together, the evidence suggests that successful quantization should not be judged only by whether a model gives the same answer, but also by whether it preserves how that answer is computed. CEQ provides one concrete way to make this mechanism-preservation objective explicit in low-bit language-model compression.

Reproducibility Statement

To the best of our ability, we have included all details necessary to reproduce our experiments in the main text and accompanying appendix. We will release our code after the review process to preserve anonymity. We fixed random seeds to ensure full reproducibility of our experiments.

AI Use Statement

During the preparation of this work, the author(s) used ChatGPT-4o to improve the grammar, language structure, and overall readability of the manuscript.

REFERENCES

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

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.

Tim Dettmers, Ruslan Svirschevski, Vage Egiazarian, Denis Kuznedelev, Elias Frantar, Saleh Ashkboos, Alexander Borzunov, Torsten Hoefler, and Dan Alistarh. SpQR: A sparse-quantized representation for near-lossless LLM weight compression. *arXiv preprint arXiv:2306.03078*, 2023.

Zhe Ding, Su Pan, and Duowei Pan. CoQuant: Joint weight-activation subspace projection for mixed-precision LLMs. *arXiv preprint arXiv:2604.26378*, 2026.

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.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. In *International Conference on Learning Representations*, 2021.

Wei Huang, Haotong Qin, Yangdong Liu, Yawei Li, Qinshuo Liu, Xianglong Liu, Luca Benini, Michele Magno, Shiming Zhang, and Xiaojuan Qi. SliM-LLM: Salience-driven mixed-precision quantization for large language models. In *International Conference on Machine Learning*, 2025.

Sehoon Kim, Coleman Hooper, Amir Gholami, Zhen Dong, Xiuyu Li, Sheng Shen, Michael W. Mahoney, and Kurt Keutzer. SqueezeLLM: Dense-and-sparse quantization. In *International Conference on Machine Learning*, 2024.

Tom Lieberum, Matthew Rahtz, Janos Kramar, Geoffrey Irving, Rohin Shah, and Vladimir Mikulik. Does circuit analysis interpretability scale? evidence from multiple choice capabilities in chinchilla. *arXiv preprint arXiv:2307.09458*, 2023.

Ji Lin, Jiaming Tang, Haotian Tang, Shang Yang, Guangxuan Xiao, and Song Han. Awq: Activation-aware weight quantization for on-device llm compression and acceleration. *GetMobile: Mobile Comp. and Comm.*, 28(4):12–17, January 2025. ISSN 2375-0529. doi: 10.1145/3714983.3714987. URL https://doi.org/10.1145/3714983.3714987.

Zechun Liu, Changsheng Zhao, Igor Fedorov, Bilge Soran, Dhruv Choudhary, Raghuraman Krishnamoorthi, Vikas Chandra, Yuandong Tian, and Tijmen Blankevoort. Spinquant: Llm quantization with learned rotations. In *International Conference on Learning Representations*, volume 2025, pages 92009–92032, 2025.

Tobias Norlund, Lovisa Hagstrom, and Richard Johansson. Transferring knowledge from vision to language: How to achieve it and how to measure it? In *BlackboxNLP Workshop*, 2021.

Pouya Pezeshkpour and Estevam Hruschka. Large language models sensitivity to the order of options in multiple-choice questions. In *Findings of NAACL*, 2024.

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. In *International Conference on Learning Representations*, 2024.

Jiuding Sun, Chantal Shaib, and Byron C. Wallace. Evaluating the zero-shot robustness of instruction-tuned language models. In *International Conference on Learning Representations*, 2024.

Yuxuan Sun, Ruikang Liu, Haoli Bai, Han Bao, Kang Zhao, Yuening Li, Jiaxin Hu, Xianzhi Yu, Lu Hou, Chun Yuan, Xin Jiang, Wulong Liu, and Jun Yao. FlatQuant: Flatness matters for LLM quantization. In *International Conference on Machine Learning*, 2025.

Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. CommonsenseQA: A question answering challenge targeting commonsense knowledge. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4149–4158, 2019.

Sarah Wiegreffe, Oyvind Tafjord, Yonatan Belinkov, Hannaneh Hajishirzi, and Ashish Sabharwal. Answer, assemble, ace: Understanding how LMs answer multiple choice questions. In *International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=6NNA0MxhCH.

Guangxuan Xiao, Ji Lin, Mickael Seznec, Hao Wu, Julien Demouth, and Song Han. SmoothQuant: Accurate and efficient post-training quantization for large language models. In *International Conference on Machine Learning*, 2023.

Hanqi Xiao, Yi-Lin Sung, Elias Stengel-Eskin, and Mohit Bansal. Task-circuit quantization: Leveraging knowledge localization and interpretability for compression. In *Conference on Language Modeling*, 2025. URL https://openreview.net/forum?id=a201nfn3xX.

Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. HellaSwag: Can a machine really finish your sentence? In *Annual Meeting of the Association for Computational Linguistics*, 2019.

Chujie Zheng, Hao Zhou, Fandong Meng, Jie Zhou, and Minlie Huang. Large language models are not robust multiple choice selectors. In *International Conference on Learning Representations*, 2024.