
OUTPUT-AWARE ROTATION FOR INT2 KV-CACHE QUANTIZATION

Vincent-Daniel Yun^{1,†}, Woosang Lim^{2,†}, Minsoo Cheong², Sunwoo Lee³
 Murali Annavaram¹, Sai Praneeth Karimireddy¹, Sungjoo Yoo^{2*}

¹University of Southern California
 {yunjuyou, annavara, karimire}@usc.edu

²Seoul National University
 {ftyg656512, icycle0409}@snu.ac.kr

³Inha University
 {sunwool}@inha.ac.kr

†Equal Contribution

ABSTRACT

The key-value (KV) cache has become a major memory and bandwidth bottleneck in long-context large language model inference, making ultra-low-bit quantization increasingly important. However, existing rotation-based INT2 methods optimize cache statistics or proxy errors before the complete attention readout, even though the model is ultimately affected by the error propagated through attention and the output projection W_O . To address this mismatch, we propose *OptR*, an output-aware rotation method that minimizes post- W_O attention-output error. OptR decomposes the post- W_O attention-output error into key- and value-induced terms and learns per-head orthogonal corrections through the full INT2 quantization and attention path. OptR further applies an attention-equivalent key reparameterization to reduce large channel-wise offsets without changing the softmax distribution. Across three models and five reasoning and coding benchmarks, OptR consistently improves both QuaRot and OSCAR and strengthens long-context retrieval, while preserving the paged KV-cache format with negligible inference overhead.



1 Introduction

As large language models (LLMs) grow in model size and context length, the key-value (KV) cache becomes a major bottleneck in long-context inference Dao et al. [2022], Ainslie et al. [2023]. During autoregressive decoding, each layer stores the keys and values of all previous tokens and reads them at every generation step. As a result, KV-cache storage and memory traffic increase with context length, batch size, and model depth. KV-cache quantization reduces these costs by storing the cache at lower precision. We focus on INT2 because it requires only 1/8 of the BF16 KV-cache storage and 1/2 of that of INT4, enabling longer contexts or larger batches under the same memory budget. Since all values in a quantization group share one scale, a few large values can expand the range represented by only four INT2 levels. This increases rounding error for most values, while aggressive clipping introduces large errors in the outliers themselves Zhang et al. [2023], Hooper et al. [2024], Zhou et al. [2026].

Rotation-based methods reduce this error by spreading a few extreme channel values across dimensions, and make the cache easier to quantize. Given an orthogonal matrix R , a cache vector z is transformed to zR before quantization and mapped back with $R^\top$ after dequantization Chee et al. [2023], Ashkboos et al. [2024]. Rotation preserves the cache shape and regular memory layout, maintaining compatibility with paged KV-cache systems and fused decoding kernels Kwon et al. [2023], Zheng et al. [2024], Zhou et al. [2026]. The main challenge is selecting R . Existing

*Corresponding Author: sungjoo.yoo@gmail.com.

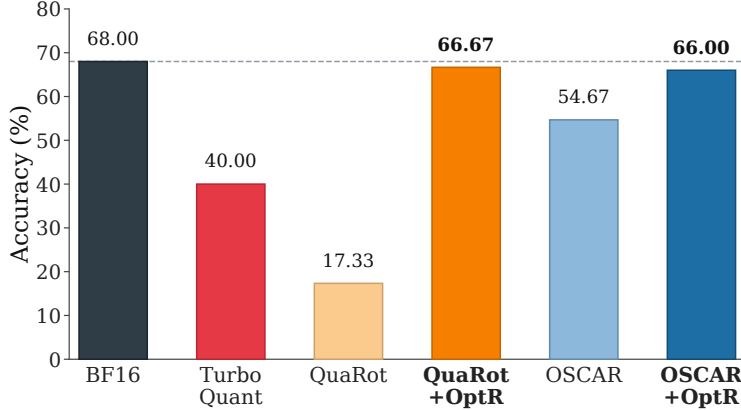


Figure 1: AIME25 accuracy of Qwen3-8B under BF16 and INT2 KV-cache quantization. + OptR denotes applying key reparameterization and output-aware rotation correction to the corresponding base rotation. The dashed line marks the BF16 accuracy.

Rotation-based INT2 KV cache pipelines rely on fixed transforms, or proxy objectives defined before the complete attention readout Ashkboos et al. [2024], Su et al. [2025], Zhou et al. [2026].

However, these proxy objectives do not directly measure the error passed to later layers. KV quantization changes the attention readout, and the output projection W_O maps this change into the model hidden space. The resulting output error enters the residual stream and propagates through subsequent layers, potentially affecting the final prediction. Consequently, the rotation that best reconstructs the cached keys and values may not be the one that best preserves the post- W_O output. This objective mismatch motivates optimizing rotations in output space.

To address this mismatch, we propose *OptR*, an output-aware rotation method for INT2 KV-cache quantization. OptR first centers the keys before rotation and quantization. This shifts all logits for a query by the same constant and therefore leaves the softmax distribution unchanged. Since INT2 has only four quantization levels, outliers can lead to large quantization errors. This reparameterization reduces their effect by narrowing the quantization range. OptR then learns per-head orthogonal corrections to any base rotation by minimizing the post- W_O attention-output error through the INT2 attention path. It optimizes the key rotation before the value rotation because the quantized keys determine the attention weights used for value aggregation. Only the rotation parameters are optimized on calibration data. Model weights remain frozen, and the learned rotations are fixed during inference.

We integrate OptR into an SGLang-based INT2 KV cache pipeline while retaining paged and prefix-cache support with negligible runtime overhead Zhou et al. [2026]. Figure 1 shows that OptR improves AIME25 accuracy on Qwen3-8B from 17.33% to 66.67% with QuaRot and from 54.67% to 66.00% with OSCAR, compared with 68.00% for BF16. The gains with both base rotations show that OptR does not depend on a specific rotation initialization.

Our contributions are summarized as follows.

- We formulate INT2 KV-cache quantization as an output-space optimization problem and decompose the post- W_O attention-output error into key- and value-induced terms.
- We propose *OptR*, which first applies attention-equivalent key reparameterization to reduce large channel-wise offsets and then learns per-head orthogonal corrections through the complete INT2 quantization and attention path.
- We show that OptR consistently improves existing rotation-based INT2 KV-cache pipelines while retaining their cache layout with negligible serving overhead.

2 Related Works

KV-cache quantization. The KV cache grows with context length and is repeatedly read during decoding, making it a major memory and bandwidth bottleneck. Prior work reduces this cost through fine-grained quantization, mixed precision, and vector quantization Liu et al. [2024], Hooper et al. [2024], Xia et al. [2026], Son et al. [2026]. These designs often introduce residual buffers, channel-wise metadata, promoted high-precision channels, or specialized cache layouts, which complicate their integration with paged KV-cache systems and fused decoding kernels. In contrast,

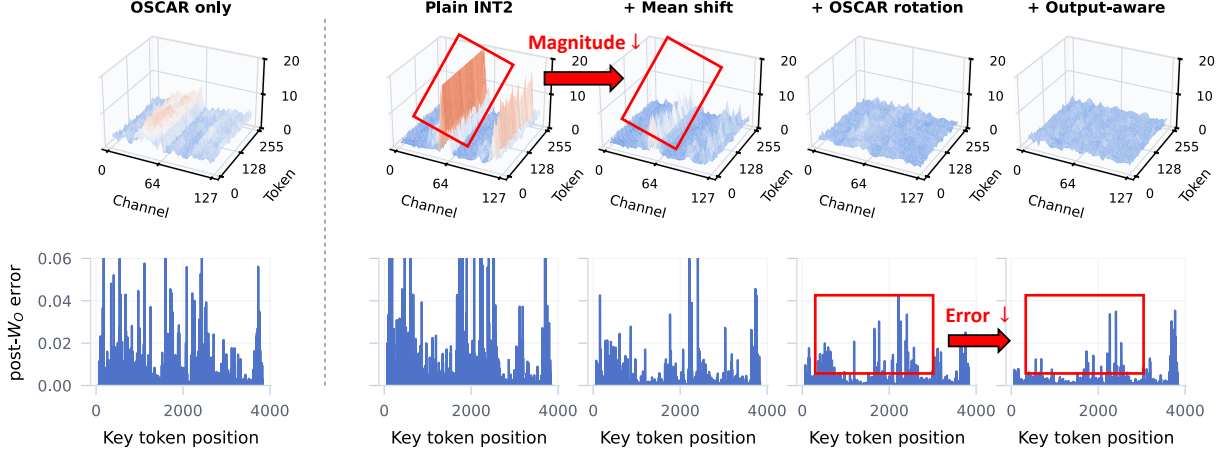


Figure 2: Key magnitude (top) and key-induced attention-output error by cached token (bottom) across five INT2 KV-cache settings on Qwen3-8B (AIME25): OSCAR only, plain INT2, reparameterization (mean shift), reparameterization with OSCAR rotation, and the full OptR pipeline. Lower is better; example details are provided in the Appendix.

rotation-based quantization transforms cached vectors into a quantization-friendly basis without changing their tensor shape or regular cache layout, making it easier to deploy in existing inference systems Ashkboos et al. [2024], Zhou et al. [2026].

Rotation-based KV-cache quantization. QuaRot uses Hadamard rotations for weights, activations, and KV caches Ashkboos et al. [2024], while RotateKV adapts rotations to head-specific key outliers and protects attention sinks Su et al. [2025]. OSCAR derives key and value rotations from offline attention-aware covariance statistics Zhou et al. [2026]. Despite these differences, existing methods select rotations using fixed transforms, cache statistics, or proxy objectives defined before the complete attention readout. Instead, OptR optimizes rotations against the post- W_O attention-output error produced by the complete INT2 attention path.

3 Problem Formulation

Preliminaries. We consider a decoder-only Transformer layer ℓ with grouped-query attention (GQA) Vaswani et al. [2017], Ainslie et al. [2023]. Let $h \in \{1, \dots, H_{kv}\}$ denote a KV head, and let $G_h \subseteq \{1, \dots, H_q\}$ denote the set of query heads that share this KV head. For each query head $j \in G_h$, we write $q_{t,j}^\ell, k_{s,h}^\ell, v_{s,h}^\ell \in \mathbb{R}^{1 \times d}$, where $s \leq t$, t is the current decoding position, and s indexes a cached source token. Equivalently, the cached keys and values for KV head h up to position t are $K_{1:t,h}^\ell = [k_{1,h}^\ell; \dots; k_{t,h}^\ell] \in \mathbb{R}^{t \times d}$, and $V_{1:t,h}^\ell = [v_{1,h}^\ell; \dots; v_{t,h}^\ell] \in \mathbb{R}^{t \times d}$. The BF16 attention logits and probabilities are:

$$a_{t,s}^{\ell,j,h} = \frac{\langle q_{t,j}^\ell, k_{s,h}^\ell \rangle}{\sqrt{d}}, \quad p_t^{\ell,j,h} = \text{softmax}_{s \leq t} \left(a_{t,s}^{\ell,j,h} \right) \quad (1)$$

The corresponding attention output is

$$o_{t,j}^\ell = \sum_{s \leq t} p_{t,s}^{\ell,j,h} v_{s,h}^\ell = (p_t^{\ell,j,h})^\top V_{1:t,h}^\ell \in \mathbb{R}^{1 \times d} \quad (2)$$

Let $W_{O,j}^\ell \in \mathbb{R}^{d_{\text{model}} \times d}$ be the output projection for query head j . The attention-output contribution of this head is:

$$y_{t,j}^\ell = \left((p_t^{\ell,j,h})^\top V_{1:t,h}^\ell \right) (W_{O,j}^\ell)^\top \in \mathbb{R}^{1 \times d_{\text{model}}} \quad (3)$$

Thus, the KV cache affects the model through the attention-weighted readout $(p_t^{\ell,j,h})^\top V_{1:t,h}^\ell$ and its projection by $W_{O,j}^\ell$.

3.1 Rotated INT2 KV-Cache Quantization

We consider long-context decoding where the long-history KV cache is stored in INT2 and a small BF16 window is preserved. Let $Q_2(\cdot; c, G)$ denote the INT2 quantize-dequantize map with clipping ratio c and group size G . For an orthogonal rotation $R \in O(d)$, define

$$D_{R,c}(z) = Q_2(zR; c, G)R^\top \quad (4)$$

If Q_2 is replaced by the identity map, then $D_{R,c}(z) = z$. Thus, the rotation only changes the coordinate system in which INT2 quantization error is introduced. We denote by $\tilde{k}_{s,h}^\ell$ and $\tilde{v}_{s,h}^\ell$ the effective keys and values used by attention after rotated INT2 quantization and BF16 window restoration.

3.2 Output-Space Error Induced by KV Quantization

We now trace the effective INT2 cache through attention and W_O and decompose the resulting post- W_O attention-output error into key- and value-induced terms. With INT2 keys, the attention logits and probabilities become

$$\tilde{a}_{t,s}^{\ell,j,h} = \frac{\langle q_{t,j}^\ell, \tilde{k}_{s,h}^\ell \rangle}{\sqrt{d}}, \quad \tilde{p}_t^{\ell,j,h} = \text{softmax}_{s \leq t} \left(\tilde{a}_{t,s}^{\ell,j,h} \right) \quad (5)$$

Let $\Delta k_{s,h}^\ell = \tilde{k}_{s,h}^\ell - k_{s,h}^\ell$ and $\Delta v_{s,h}^\ell = \tilde{v}_{s,h}^\ell - v_{s,h}^\ell$ denote the key and value quantization errors. Key errors first perturb the attention logits:

$$\Delta a_{t,s}^{\ell,j,h} = \tilde{a}_{t,s}^{\ell,j,h} - a_{t,s}^{\ell,j,h} = \frac{\langle q_{t,j}^\ell, \Delta k_{s,h}^\ell \rangle}{\sqrt{d}} \quad (6)$$

The resulting attention error is $\Delta p_t^{\ell,j,h} = \tilde{p}_t^{\ell,j,h} - p_t^{\ell,j,h}$.

Thus, the effect of a key error depends on the query and the softmax attention map, not only on $\|\Delta k\|_2^2$.

Under INT2 keys and values, the attention-output contribution becomes

$$\tilde{y}_{t,j}^\ell = \left(\sum_{s \leq t} \tilde{p}_{t,s}^{\ell,j,h} \tilde{v}_{s,h}^\ell \right) (W_{O,j}^\ell)^\top \quad (7)$$

Subtracting the BF16 readout gives the exact decomposition. Let $\Delta y_{t,j}^\ell := \tilde{y}_{t,j}^\ell - y_{t,j}^\ell$. Then

$$\Delta y_{t,j}^\ell = \underbrace{\left(\sum_{s \leq t} \Delta p_{t,s}^{\ell,j,h} v_{s,h}^\ell \right) (W_{O,j}^\ell)^\top}_{\text{key-induced output error}} + \underbrace{\left(\sum_{s \leq t} \tilde{p}_{t,s}^{\ell,j,h} \Delta v_{s,h}^\ell \right) (W_{O,j}^\ell)^\top}_{\text{value-induced output error}} \quad (8)$$

We denote the two terms above by $\delta y_{K,t,j}^{\ell,h}$ and $\delta y_{V,t,j}^{\ell,h}$, respectively. Here $\delta y_{K,t,j}^{\ell,h}$ is the output error induced by key quantization through the attention distribution, while $\delta y_{V,t,j}^{\ell,h}$ is the value error after attention-weighted aggregation and output projection.

This decomposition shows why raw cache reconstruction is only a proxy. A reconstruction-based objective measures

$$E_{\text{rec}} = \|K - \tilde{K}\|_F^2 + \|V - \tilde{V}\|_F^2 \quad (9)$$

whereas the model observes the attention-output error

$$E_{\text{out}} = \|\tilde{y}_{t,j}^\ell - y_{t,j}^\ell\|_2^2 = \left\| \delta y_{K,t,j}^{\ell,h} + \delta y_{V,t,j}^{\ell,h} \right\|_2^2 \quad (10)$$

Eq. (9) and Eq. (10) can favor different rotations because attention and W_O reduce the effect of some cache errors while allowing others to affect the attention-output. For this reason, OptR uses E_{out} as its rotation optimization target.

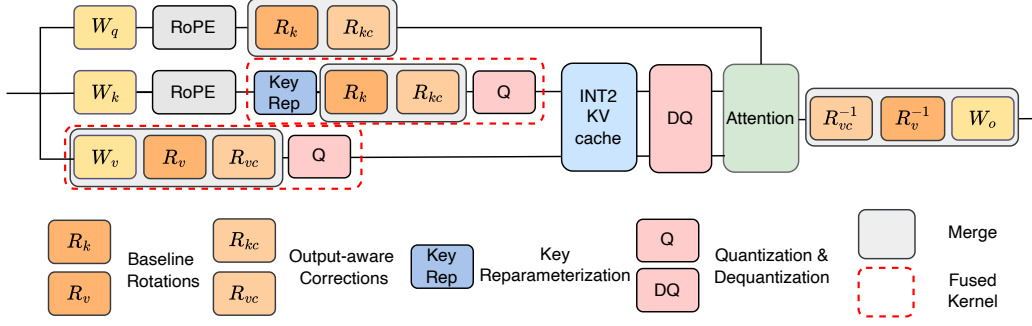


Figure 3: Overview of OptR. OptR augments an existing rotation-based KV-cache quantization pipeline with channel-wise key reparameterization and output-aware correction rotations before INT2 quantization. The resulting quantized KV cache is consumed by attention and projected through W_O , where OptR targets reduced post- W_O attention-output error.

4 Method: Output-Aware Rotation

Figure 5 shows that INT2-induced output error differs across KV heads. OptR therefore learns a separate orthogonal correction to the base key and value rotations for each head. During one-time offline calibration, model weights remain frozen and only the corrections are optimized to reduce post- W_O attention-output error through the INT2 attention path. We optimize the key rotation first because the INT2 keys determine the attention distribution, and then optimize the value rotation under this distribution. The resulting rotations are fixed during inference. Figure 3 summarizes the full pipeline. We omit (ℓ, h) when the layer and KV head are clear.

4.1 Rotated INT2 Cache Reparameterization

As shown in Figure 2, large channel-wise key offsets increase the dynamic range of group-wise INT2 quantization. Therefore, we reparameterize keys using the per-channel calibration mean $\mu \in \mathbb{R}^d$ before rotation and quantization. For key and value rotations R_K and R_V , OptR defines

$$\bar{k}_s(R_K) = D_{R_K, c_K}(k_s - \mu), \quad \bar{v}_s(R_V) = D_{R_V, c_V}(v_s)$$

This key reparameterization is attention-equivalent: subtracting the same μ from every key adds a query-dependent constant to all logits and leaves the softmax distribution unchanged. We apply it only to keys, since shifting values would alter the attention output. Sink and recent tokens remain in BF16, and the same centering is applied to their keys. The effective cache is denoted by $\bar{k}_s(R_K)$ and $\bar{v}_s(R_V)$.

4.2 Orthogonal Rotation Refinement

OptR does not rely on a specific rotation initialization. Let $R_K^0, R_V^0 \in O(d)$ denote arbitrary orthogonal initializations for the key and value rotations, respectively. Given these initial rotations, OptR applies the same output-aware rotation procedure regardless of how they are constructed.

For each KV head, OptR learns unconstrained matrices $A_K, A_V \in \mathbb{R}^{d \times d}$ and forms the skew-symmetric generators

$$S_K = A_K - A_K^\top, \quad S_V = A_V - A_V^\top \quad (11)$$

which define the corrected rotations

$$R_K(A_K) = R_K^0 \exp(S_K), \quad R_V(A_V) = R_V^0 \exp(S_V)$$

Since S_K and S_V are skew-symmetric, their matrix exponentials are orthogonal. Therefore, the corrected rotations remain orthogonal throughout calibration. The same formulation applies to different rotation initializations without modifying the calibration objective.

4.3 Output-Aware Key Calibration

The key rotation affects attention logits and the attention distribution next. To isolate this path, we calibrate the key rotation while keeping values in BF16. The induced key-only attention distribution is

$$a_{t,s}^K(R_K) = \frac{\langle q_{t,j}, \tilde{k}_s(R_K) \rangle}{\sqrt{d}}, \quad p_t^K(R_K) = \text{softmax}_{s \leq t} (a_{t,s}^K(R_K)) \quad (12)$$

The key-induced attention-output error is

$$e_K(t, j; R_K) = \left[\sum_{s \leq t} (p_{t,s}^K(R_K) - p_{t,s}) v_s \right] (W_{O,j})^\top \quad (13)$$

We optimize the key rotation with

$$L_K(R_K; D) = \mathbb{E}_{(t,j) \in D} \left[D_{\text{KL}}(p_t \| p_t^K(R_K)) + \lambda_K \frac{\|e_K(t, j; R_K)\|_2^2}{d_{\text{model}}} \right] \quad (14)$$

The KL term preserves the attention distribution, while the second term measures the post- W_O attention-output error caused by key-induced attention changes.

4.4 Output-Aware Value Calibration

After optimizing the key rotation, we calibrate the value rotation under the selected quantized-key attention path. Let $\hat{R}_K$ be the selected key rotation and let $\hat{p}_t^K$ be its induced attention distribution.

The value-induced attention-output error is

$$e_V(t, j; R_V) = \left[\sum_{s \leq t} \hat{p}_{t,s}^K (\tilde{v}_s(R_V) - v_s) \right] (W_{O,j})^\top \quad (15)$$

We optimize the value rotation with

$$L_V(R_V; D) = \mathbb{E}_{(t,j) \in D} \left[\frac{\|e_V(t, j; R_V)\|_2^2}{d_{\text{model}}} \right] \quad (16)$$

This objective measures value-induced attention-output error after attention-weighted aggregation and output projection, rather than raw value-cache reconstruction.

4.5 INT2-Aware Orthogonal Calibration

OptR calibrates the rotation generators through the full INT2 quantization path: rotation, clipping, grouping, INT2 rounding, dequantization, and inverse rotation. For each layer and KV head, OptR learns the orthogonal correction:

$$A_K^* = \arg \min_A L_K(R_K^0 \exp(A - A^\top); D_{\text{train}}), \quad A_V^* = \arg \min_A L_V(R_V^0 \exp(A - A^\top); D_{\text{train}}) \quad (17)$$

All model weights remain frozen; only the per-head rotation generators are calibrated. Since INT2 rounding is non-differentiable, gradients are estimated using a straight-through estimator:

$$\frac{\partial Q_2(Z; c, G)}{\partial Z} \approx \mathbf{1}\{|Z| \leq \tau_c\} \quad (18)$$

where τ_c is the clipping threshold determined by the clip ratio c .

5 Experimental Results

Models and Benchmarks. We evaluate OptR on Qwen3-4B-Thinking-2507 Yang et al. [2025], Qwen3-8B Yang et al. [2025], and Phi4-14B-reasoning-plus Abdin et al. [2025]. We measure reasoning and coding accuracy on AIME24, AIME25 MAA [2025], GPQA-Diamond Rein et al. [2024], MBPP+ Liu et al. [2023], and LiveCodeBench v6 Jain et al. [2025]. We additionally evaluate long-context retrieval using RULER-NIAH (Needle-in-a-Haystack) Hsieh et al. [2024], with context lengths up to 64K for the Qwen3 models and 32K for Phi4-14B-reasoning-plus. For all models, we use a temperature of 0.6, top- p of 0.95, and top- k of 20. For the reasoning and coding benchmarks, we use maximum generation lengths of 32K for the Qwen3 models and 16K for Phi4-14B-reasoning-plus.

Model	Method	BPE	AIME24	AIME25	GPQA	MBPP+	LCB v6	Mean
Qwen3-4B-Thinking-2507	BF16	16	78.00 $\pm$ 6.91	71.33 $\pm$ 3.80	64.55 $\pm$ 2.30	77.31 $\pm$ 0.87	46.71 $\pm$ 2.05	67.58
	TurboQuant (no MP)	3.25	10.00	16.67	41.92	23.02	1.71	18.66
	QuaRot-INT2	2.32	0.00 $\pm$ 0.00	0.00 $\pm$ 0.00	6.06 $\pm$ 1.18	5.61 $\pm$ 2.36	1.37 $\pm$ 0.31	2.61
	QuaRot-INT2 + OptR	2.32	69.33 $\pm$ 4.35	60.67 $\pm$ 4.35	62.12 $\pm$ 2.92	78.31 $\pm$ 1.23	37.71 $\pm$ 0.81	61.63
	OSCAR	2.32	68.67 $\pm$ 2.98	63.33 $\pm$ 4.08	62.02 $\pm$ 1.73	75.66 $\pm$ 0.84	43.89 $\pm$ 0.75	62.71
	OSCAR + OptR	2.32	72.00 $\pm$ 1.83	70.67 $\pm$ 2.79	62.83 $\pm$ 2.22	76.83 $\pm$ 0.91	44.57 $\pm$ 0.90	65.38
Qwen3-8B	BF16	16	76.00 $\pm$ 6.41	68.00 $\pm$ 2.98	55.76 $\pm$ 2.16	79.15 $\pm$ 1.48	49.37 $\pm$ 0.65	65.66
	TurboQuant (no MP)	3.25	53.33	40.00	45.96	58.99	21.14	43.88
	QuaRot-INT2	2.32	16.00 $\pm$ 5.48	17.33 $\pm$ 4.35	42.12 $\pm$ 1.54	49.10 $\pm$ 1.00	5.26 $\pm$ 1.24	25.96
	QuaRot-INT2 + OptR	2.32	76.00 $\pm$ 4.35	66.67 $\pm$ 3.33	57.58 $\pm$ 2.45	80.63 $\pm$ 1.14	39.77 $\pm$ 1.54	64.13
	OSCAR	2.32	72.00 $\pm$ 3.80	54.67 $\pm$ 3.80	56.26 $\pm$ 3.25	78.73 $\pm$ 0.44	45.03 $\pm$ 2.26	61.34
	OSCAR + OptR	2.32	76.67 $\pm$ 2.36	66.00 $\pm$ 4.94	58.99 $\pm$ 2.56	79.21 $\pm$ 0.55	47.54 $\pm$ 0.48	65.68
Phi4-14B-reasoning-plus	BF16	16	70.00 $\pm$ 4.71	60.67 $\pm$ 4.94	45.56 $\pm$ 2.09	77.18 $\pm$ 2.16	39.43 $\pm$ 2.06	58.57
	TurboQuant (no MP)	3.25	60.00	53.33	46.97	74.60	33.14	53.61
	QuaRot-INT2	2.32	61.33 $\pm$ 4.47	46.00 $\pm$ 5.48	46.77 $\pm$ 3.40	75.82 $\pm$ 1.16	34.97 $\pm$ 1.48	52.98
	QuaRot-INT2 + OptR	2.32	64.00 $\pm$ 3.65	52.00 $\pm$ 2.98	47.58 $\pm$ 2.91	76.70 $\pm$ 0.67	36.00 $\pm$ 0.57	55.26
	OSCAR	2.32	62.67 $\pm$ 2.79	49.33 $\pm$ 3.65	43.84 $\pm$ 2.71	73.02 $\pm$ 1.44	34.29 $\pm$ 1.04	52.63
	OSCAR + OptR	2.32	65.33 $\pm$ 5.06	58.00 $\pm$ 5.58	46.16 $\pm$ 1.50	73.44 $\pm$ 0.55	35.09 $\pm$ 1.11	55.60

Table 1: Comparison of INT2 KV-cache quantization methods across three model configurations and five benchmarks. Results are reported as $\mu \pm \sigma$ over five seeds. OptR indicates that output-aware rotation is applied to the corresponding baseline. BPE denotes the effective number of bits per KV-cache element. TurboQuant is reported from a single run because repeated 32K evaluations are slow in its vLLM implementation.

Model	Method	4k	8k	16k	32k	64k
Qwen3-4B-Thinking-2507	BF16	99.90 $\pm$ 0.08	99.60 $\pm$ 0.07	98.19 $\pm$ 0.08	97.13 $\pm$ 0.11	88.76 $\pm$ 0.47
	QuaRot-INT2	3.08 $\pm$ 0.19	8.26 $\pm$ 0.85	0.00 $\pm$ 0.00	6.14 $\pm$ 1.11	0.72 $\pm$ 0.16
	QuaRot-INT2 + OptR	99.78 $\pm$ 0.17	99.02 $\pm$ 0.38	95.59 $\pm$ 0.52	74.69 $\pm$ 0.99	46.86 $\pm$ 1.77
	OSCAR	99.29 $\pm$ 0.31	97.59 $\pm$ 0.50	93.78 $\pm$ 0.33	77.75 $\pm$ 0.64	55.72 $\pm$ 1.44
	OSCAR + OptR	99.42 $\pm$ 0.27	98.57 $\pm$ 0.17	95.31 $\pm$ 0.10	84.98 $\pm$ 0.45	69.58 $\pm$ 1.31
Qwen3-8B	BF16	99.83 $\pm$ 0.11	99.93 $\pm$ 0.02	99.45 $\pm$ 0.08	98.70 $\pm$ 0.49	84.22 $\pm$ 1.41
	QuaRot-INT2	84.97 $\pm$ 0.09	41.73 $\pm$ 5.75	18.16 $\pm$ 1.14	13.03 $\pm$ 1.71	0.04 $\pm$ 0.07
	QuaRot-INT2 + OptR	99.40 $\pm$ 0.29	98.76 $\pm$ 0.57	96.14 $\pm$ 0.35	86.37 $\pm$ 0.72	70.02 $\pm$ 1.79
	OSCAR	99.59 $\pm$ 0.15	97.94 $\pm$ 0.28	94.39 $\pm$ 0.29	83.76 $\pm$ 0.58	57.54 $\pm$ 1.82
	OSCAR + OptR	99.60 $\pm$ 0.11	98.16 $\pm$ 0.63	95.50 $\pm$ 0.65	86.42 $\pm$ 0.55	68.65 $\pm$ 0.80
Phi4-14B-reasoning-plus	BF16	98.50 $\pm$ 0.45	97.54 $\pm$ 0.59	97.76 $\pm$ 0.70	92.13 $\pm$ 1.34	N/A
	QuaRot-INT2	96.28 $\pm$ 0.44	92.28 $\pm$ 1.31	84.54 $\pm$ 0.52	69.65 $\pm$ 0.94	N/A
	QuaRot-INT2 + OptR	97.62 $\pm$ 0.21	94.02 $\pm$ 0.80	89.98 $\pm$ 0.52	76.21 $\pm$ 0.53	N/A
	OSCAR	95.62 $\pm$ 1.02	90.50 $\pm$ 0.24	84.59 $\pm$ 0.86	70.09 $\pm$ 0.59	N/A
	OSCAR + OptR	96.30 $\pm$ 0.37	93.42 $\pm$ 0.35	87.64 $\pm$ 0.45	75.36 $\pm$ 0.85	N/A

Table 2: RULER-NIAH long-context retrieval accuracy evaluated at context lengths ranging from 4k to 64k tokens. Results are reported as $\mu \pm \sigma$ over three random seeds, with 800 examples evaluated per seed. OptR indicates that output-aware rotation is applied to the corresponding baseline. Since Phi4-14B-reasoning-plus supports a maximum context length of 32k, its evaluation is limited to 32k, and the 64k setting is reported as N/A. Additional 128K results are in the Appendix.

Implementation Details. OptR is implemented on top of the official OSCAR codebase Zhou et al. [2026]. We use group-wise affine INT2 quantization with a group size of 128, clipping ratios of 0.96 for keys and 0.92 for values, and retain 64 sink tokens and 256 recent tokens in BF16. λ_K is fixed to 1.0 across all models and benchmarks. We report an effective cache cost of 2.32 bits per element (BPE) at a 64K-token context, including quantization metadata and the BF16 windows. For each model, we collect a 30K-token pool of BF16 GPQA QKV traces and use disjoint subsets for calibration and held-out rotation selection. For each layer and KV head, we optimize only the rotation corrections for 80 Adam steps with a learning rate of 0.02 Kingma and Ba [2014]. All model weights remain frozen. All experiments were conducted on four NVIDIA A100 40GB GPUs. Additional implementation details are provided in the Appendix.

Baselines. We compare OptR with BF16, TurboQuant Zandieh et al. [2026] without mixed precision (no MP), QuaRot-INT2 Ashkboos et al. [2024], and OSCAR Zhou et al. [2026]. OptR is applied to both QuaRot and OSCAR to evaluate its effectiveness across fixed Hadamard rotations and the state-of-the-art attention-aware covariance rotation. All rotation-based methods use the same INT2 group size and BF16 sink and recent-token windows.

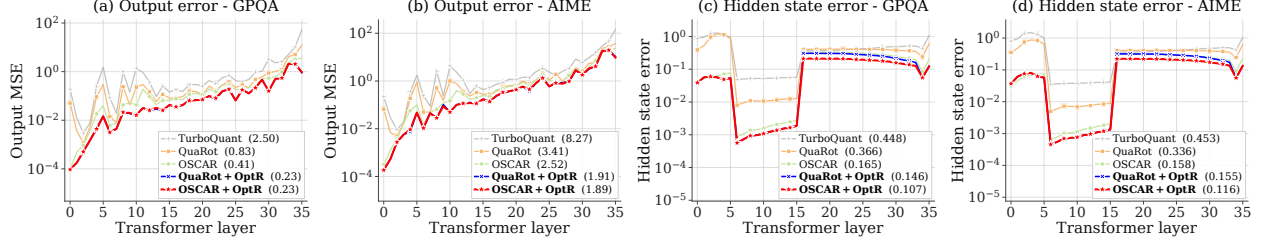


Figure 4: Per-layer errors under INT2 KV-cache quantization on Qwen3-4B-Thinking-2507. Panels (a,b) show post- W_O attention-output error on GPQA-Diamond and AIME25 at each layer. Panels (c,d) show the propagated residual-stream error, measured as the normalized squared error between BF16 and INT2-KV block-output hidden states on the same BF16-generated token sequences. Legend values indicate the mean across layers. Error values are shown on a logarithmic scale. Detailed settings are provided in Appendix.

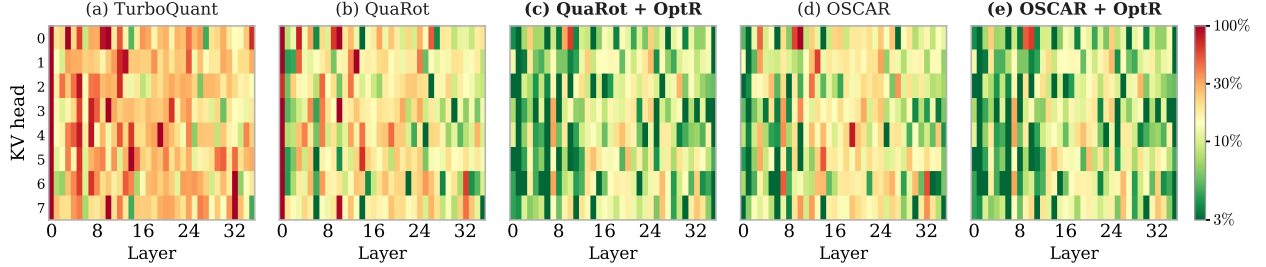


Figure 5: Output error by layer and KV head relative to naive INT2 (Qwen3-4B-Thinking-2507, GPQA-Diamond). Each cell reports the post- W_O attention-output RMS error as a percentage of plain per-group INT2 without rotation. Lower values indicate smaller output error, with naive INT2 corresponding to 100%. Panels show (a) TurboQuant, (b) QuaRot, (c) QuaRot + OptR, (d) OSCAR, and (e) OSCAR + OptR. Detailed settings are provided in Appendix.

5.1 Main Results

Overall Accuracy. Table 1 compares OptR with BF16 and existing INT2 KV-cache quantization methods. QuaRot-INT2 often suffers from severe accuracy degradation, particularly on reasoning and coding tasks. Applying OptR substantially improves both QuaRot and OSCAR at the same 2.32 BPE across the evaluated models and benchmarks.

Long-Context Robustness. Table 2 reports RULER-NIAH Hsieh et al. [2024] retrieval accuracy across increasing context lengths. QuaRot-INT2 degrades rapidly as the context grows, whereas OptR preserves substantially stronger retrieval performance for both QuaRot and OSCAR. The gains become more pronounced at longer contexts, indicating that OptR more effectively mitigates accumulated KV quantization error over long histories.

Output-Error Analysis. Figure 4 shows that OptR reduces both post- W_O attention-output error and propagated residual-stream error relative to QuaRot and OSCAR on GPQA-Diamond and AIME25. Figure 5 further shows that these reductions occur across most layers and KV heads rather than only a small subset.

Ablation Study. All ablation results are reported as $\mu \pm \sigma$ over five seeds. Table 4 compares cache-reconstruction, pre- W_O , and post- W_O calibration objectives on Qwen3-4B-Thinking-2507 using the same OSCAR base rotation. The post- W_O attention-output objective achieves the highest AIME25 accuracy. Table 3 shows that the full OptR consistently achieves the strongest performance across both base rotations and models, with key reparameterization providing an additional source of improvement. Additional sensitivity ablation results for λ_K tuning are provided in the Appendix.

6 Discussion

Efficiency Analysis. We integrate OptR into an optimized INT2 KV-cache pipeline and extend its Triton cache-write kernel to support per-KV-head key rotations and key reparameterization. For each head h , reparameterization before rotation can be rewritten as

$$(k - \mu_h)R_{K,h} = kR_{K,h} - \mu_h R_{K,h}$$

Method	Qwen3-4B-Thinking-2507	Phi4-14B-reasoning-plus
BF16	71.33 $\pm$ 3.80	60.67 $\pm$ 4.94
OSCAR	63.33 $\pm$ 4.08	49.33 $\pm$ 3.65
OSCAR + Key Reparam	66.00 $\pm$ 4.71	54.67 $\pm$ 6.91
OSCAR + OptR	70.67 $\pm$ 2.79	58.00 $\pm$ 5.58

Table 3: Component-wise evaluation of OptR on AIME25 across Qwen3-4B-Thinking-2507 and Phi4-14B-reasoning-plus using the state-of-the-art OSCAR rotation.

Key Objective $\mathcal{L}_K$	Value Objective $\mathcal{L}_V$	AIME25
$D_{KL} + \lambda_K \mathcal{E}(\tilde{K} - K)$	$\mathcal{E}(\tilde{V} - V)$	65.33 $\pm$ 6.91
$D_{KL} + \lambda_K \mathcal{E}(\Delta p_K^\top V)$	$\mathcal{E}((\hat{p}_t^K)^\top \Delta V)$	62.67 $\pm$ 2.79
$D_{KL} + \lambda_K \mathcal{E}((\Delta p_K^\top V) W_{O,j}^\top)$	$\mathcal{E}((\hat{p}_t^K)^\top \Delta V) W_{O,j}^\top)$	70.67 $\pm$ 2.79

Table 4: Objective ablation for OptR on Qwen3-4B-Thinking-2507 with the OSCAR base rotation. $\mathcal{E}(X) = \text{mean}(X^2)$ denotes the element-wise mean squared error. The D_{KL} term is fixed across all rows, and the three rows vary the error term using cache reconstruction, pre- W_O attention readout, and post- W_O attention-output error, respectively.

We therefore precompute the rotated mean $\mu_h R_{K,h}$ once offline and subtract it immediately after rotating each key. The same kernel then clips, quantizes, packs, and writes the reparameterized keys and rotated values to the INT2 cache in a single launch. The value rotation and its inverse are absorbed into the value projection and W_O , respectively, avoiding separate value-side rotation kernels. To match the per-KV-head key rotations under GQA, we implement an optimized Triton kernel that applies the corresponding rotation to each query head.

Figure 6 shows that OptR remains within 2% of the corresponding base pipeline in decode latency, end-to-end throughput, and prefill time across the evaluated settings, while retaining the same maximum batch sizes. For the 2K-token input and 4K-token output workload in Figure 6(b), adding OptR at $B = 128$ increases GPU memory by only 18 MiB, from 36,205 MiB to 36,223 MiB. Thus, OptR preserves the efficiency and serving capacity of the underlying INT2 pipeline.

7 Conclusion

We introduced OptR, an output-aware rotation method for INT2 KV-cache quantization. OptR combines attention-equivalent key reparameterization with per-head orthogonal corrections optimized through the complete quantized attention path, directly preserving the attention output observed by subsequent layers. Across diverse models, reasoning and coding tasks, and long-context retrieval settings, OptR consistently strengthens existing rotation-based methods and reduces quantization-induced output error. Moreover, OptR retains compatibility with paged KV-cache serving and introduces negligible inference overhead, demonstrating that output-aware optimization is an effective and practical direction for ultra-low-bit KV-cache quantization.

References

- Tri Dao, Daniel Y. Fu, Stefano Ermon, Atri Rudra, and Christopher Ré. Flashattention: fast and memory-efficient exact attention with io-awareness. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, 2022.
- Joshua Ainslie, James Lee-Thorp, Michiel de Jong, Yury Zemlyanskiy, Federico Lebrón, and Sumit Sanghai. GQA: Training generalized multi-query transformer models from multi-head checkpoints. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4895–4901, 2023.
- Zhenyu Zhang, Ying Sheng, Tianyi Zhou, Tianlong Chen, Lianmin Zheng, Ruisi Cai, Zhao Song, Yuandong Tian, Christopher Ré, Clark Barrett, Zhangyang Wang, and Beidi Chen. H2o: heavy-hitter oracle for efficient generative inference of large language models. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, 2023.

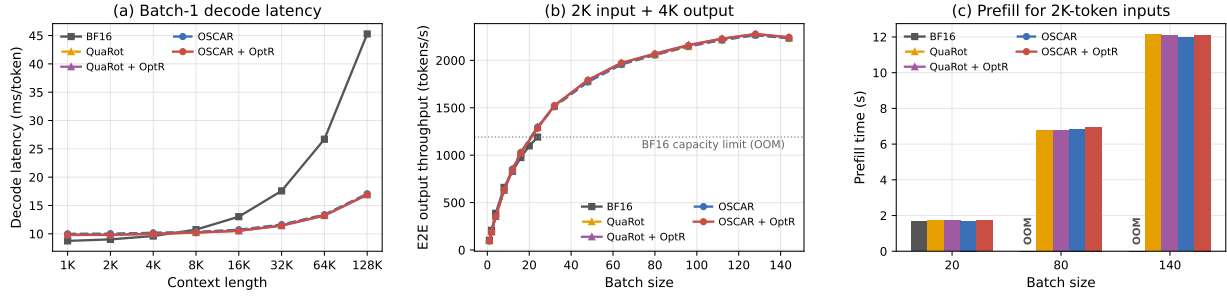


Figure 6: Efficiency analysis of OptR integrated into the optimized rotated INT2 KV cache pipeline on an NVIDIA A100 40GB using Qwen3-4B-Thinking-2507. (a) Batch-1 decode latency across context lengths from 1K to 128K. (b) End-to-end output throughput, for a 2K-input and 4K-output workload at different batch sizes. (c) Prefill time for 2K-token inputs at representative batch sizes. Analysis shows that OptR adds negligible inference overhead to the underlying pipeline.

Coleman Hooper, Sehoon Kim, Hiva Mohammadzadeh, Michael W. Mahoney, Yakun Sophia Shao, Kurt Keutzer, and Amir Gholami. Kvquant: towards 10 million context length llm inference with kv cache quantization. In *Proceedings of the 38th International Conference on Neural Information Processing Systems*, 2024.

Zhongzhu Zhou, Donglin Zhuang, Jisen Li, Ziyang Chen, Shuaiwen Leon Song, Ben Athiwaratkun, and Xiaoxia Wu. Oscar: Offline spectral covariance-aware rotation for 2-bit kv cache quantization. 2026. URL <https://arxiv.org/abs/2605.17757>.

Jerry Chee, Yaohui Cai, Volodymyr Kuleshov, and Christopher De Sa. QuIP: 2-bit quantization of large language models with guarantees. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL <https://openreview.net/forum?id=xrk9g5vcXR>.

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. *Advances in Neural Information Processing Systems*, 37:100213–100240, 2024.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph Gonzalez, Hao Zhang, and Ion Stoica. Efficient memory management for large language model serving with PagedAttention. In *Proceedings of the 29th Symposium on Operating Systems Principles (SOSP)*, 2023.

Lianmin Zheng, Liangsheng Yin, Zhiqiang Xie, Chuyue Sun, Jeff Huang, Cody Hao Yu, Shiyi Cao, Christos Kozyrakis, Ion Stoica, Joseph E. Gonzalez, Clark Barrett, and Ying Sheng. Sglang: efficient execution of structured language model programs. In *Proceedings of the 38th International Conference on Neural Information Processing Systems*, 2024.

Zunhai Su, Hanyu Wei, Zhe Chen, Wang Shen, Linge Li, Huangqi Yu, and Kehong Yuan. Rotatekv: accurate and robust 2-bit kv cache quantization for llms via outlier-aware adaptive rotations. In *Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence*, pages 6200–6208, 2025.

Zirui Liu, Jiayi Yuan, Hongye Jin, Shaochen Zhong, Zhaozhao Xu, Vladimir Braverman, Beidi Chen, and Xia Hu. Kivi: A tuning-free asymmetric 2bit quantization for kv cache. In *International Conference on Machine Learning*, pages 32332–32344. PMLR, 2024.

Haojun Xia, Xiaoxia Wu, Jisen Li, Tsai chuan Wu, Junxiong Wang, Jue WANG, Chenxi Li, Aman Singhal, Alay Dilip-hai Shah, Alpay Ariyak, Donglin Zhuang, Zhongzhu Zhou, Ben Athiwaratkun, Zhen Zheng, and Shuaiwen Leon Song. Kitty: Accurate and efficient 2-bit KV cache quantization with dynamic channel-wise precision boost. In *Ninth Conference on Machine Learning and Systems*, 2026. URL <https://openreview.net/forum?id=r3mQiuYKIN>.

Donghyun Son, Euntae Choi, and Sungjoo Yoo. Nsnquant: A double normalization approach for calibration-free low-bit vector quantization of kv cache. *Advances in Neural Information Processing Systems*, 38:43124–43159, 2026.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.

-
- Marah Abdin, Sahaj Agarwal, Ahmed Awadallah, Vidhisha Balachandran, Harkirat Behl, Lingjiao Chen, Gustavo de Rosa, Suriya Gunasekar, Mojan Javaheripi, Neel Joshi, et al. Phi-4-reasoning technical report. *arXiv preprint arXiv:2504.21318*, 2025.
- MAA. Aime 2025: American invitational mathematics examination. <https://maa.org/math-competitions/aime>, 2025.
- David Rein, Betty Li Hou, Asa Cooper Stickland, Jackson Petty, Richard Yuanzhe Pang, Julien Dirani, Julian Michael, and Samuel R. Bowman. GPQA: A graduate-level google-proof q&a benchmark. In *First Conference on Language Modeling*, 2024. URL <https://openreview.net/forum?id=Ti67584b98>.
- Jiawei Liu, Chunqiu Steven Xia, Yuyao Wang, and Lingming Zhang. Is your code generated by chatGPT really correct? rigorous evaluation of large language models for code generation. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL <https://openreview.net/forum?id=1qv610Cu7>.
- Naman Jain, King Han, Alex Gu, Wen-Ding Li, Fanjia Yan, Tianjun Zhang, Sida Wang, Armando Solar-Lezama, Koushik Sen, and Ion Stoica. Livecodebench: Holistic and contamination free evaluation of large language models for code. In *The Thirteenth International Conference on Learning Representations*, 2025. URL <https://openreview.net/forum?id=chfJJYC3iL>.
- Cheng-Ping Hsieh, Simeng Sun, Samuel Krman, Shantanu Acharya, Dima Rekesh, Fei Jia, and Boris Ginsburg. RULER: What’s the real context size of your long-context language models? In *First Conference on Language Modeling*, 2024. URL <https://openreview.net/forum?id=kIoBbc76Sy>.
- Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.
- Amir Zandieh, Majid Daliri, Majid Hadian, and Vahab Mirrokni. Turboquant: Online vector quantization with near-optimal distortion rate. In *The Fourteenth International Conference on Learning Representations*, 2026. URL <https://openreview.net/forum?id=t03ASKZlok>.

A Supplementary Materials

Additional implementation details, experimental settings, and qualitative analyses are provided in the supplementary materials.

OptR Algorithm. Algorithm 1 summarizes the one-time offline calibration of OptR. For each layer and KV head, OptR estimates the key mean and optimizes only the key and value rotation corrections while keeping all model weights frozen. The matrices A_K and A_V parameterize orthogonal corrections through $\exp(A - A^\top)$. We use $\tilde{K}$ and $\tilde{V}$ to denote the effective keys and values consumed by attention after rotated INT2 quantization. Sink and recent tokens skip INT2 quantization; their keys retain the same centering, while their values remain unchanged in BF16.

Algorithm 1 OptR offline calibration

Require: Calibration traces $\mathcal{D}$, base rotations R_K^0, R_V^0 , key-loss weight λ_K , optimization steps T

Ensure: $\{\hat{R}_{K,\ell,h}, \hat{R}_{V,\ell,h}, \mu_{\ell,h}\}_{\ell,h}$

```

1: for all layers  $\ell$  and KV heads  $h$  do
2:   Estimate the key mean  $\mu_{\ell,h}$  from  $\mathcal{D}$ 
3:   Initialize  $A_K, A_V \leftarrow 0$ 
4:   for  $i = 1, \dots, T$  do
5:      $R_K \leftarrow R_K^0 \exp(A_K - A_K^\top)$ 
6:     Construct effective keys  $\tilde{K}(R_K)$  through key reparameterization and rotated INT2 quantization
7:      $p_t^K \leftarrow \text{softmax}\left(q_{t,j} \tilde{K}_{1:t}(R_K)^\top / \sqrt{d}\right)$ 
8:      $e_K(t, j) \leftarrow [(p_t^K - p_t)^\top V_{1:t}] W_{O,j}^\top$ 
9:      $\mathcal{L}_K \leftarrow \mathbb{E}_{(t,j) \in \mathcal{D}} \left[ D_{\text{KL}}(p_t \| p_t^K) + \lambda_K \frac{\|e_K(t,j)\|_2^2}{d_{\text{model}}} \right]$ 
10:    Update  $A_K$  using Adam and the INT2 STE
11:  end for
12:   $\hat{R}_K \leftarrow R_K^0 \exp(A_K - A_K^\top)$ 
13:  Compute  $\hat{p}_t^K$  using  $\tilde{K}(\hat{R}_K)$ 
14:  for  $i = 1, \dots, T$  do
15:     $R_V \leftarrow R_V^0 \exp(A_V - A_V^\top)$ 
16:    Construct effective values  $\tilde{V}(R_V)$  through rotated INT2 quantization
17:     $e_V(t, j) \leftarrow [(\hat{p}_t^K)^\top (\tilde{V}_{1:t}(R_V) - V_{1:t})] W_{O,j}^\top$ 
18:     $\mathcal{L}_V \leftarrow \mathbb{E}_{(t,j) \in \mathcal{D}} \left[ \frac{\|e_V(t,j)\|_2^2}{d_{\text{model}}} \right]$ 
19:    Update  $A_V$  using Adam and the INT2 STE
20:  end for
21:   $\hat{R}_V \leftarrow R_V^0 \exp(A_V - A_V^\top)$ 
22:  Store  $\hat{R}_{K,\ell,h} \leftarrow \hat{R}_K$  and  $\hat{R}_{V,\ell,h} \leftarrow \hat{R}_V$ 
23: end for
24: return  $\{\hat{R}_{K,\ell,h}, \hat{R}_{V,\ell,h}, \mu_{\ell,h}\}_{\ell,h}$ 

```

Here, p_t denotes the BF16 attention distribution, p_t^K is computed using the current effective INT2 keys, and $\hat{p}_t^K$ is computed using the selected key rotation $\hat{R}_K$. The values $V_{1:t}$ remain in BF16 during key calibration. The projection $W_{O,j}$ maps the output of query head j into the model hidden space. Layer and KV-head indices are omitted inside the algorithm when clear.

Details of Figure 2. We use the same 4,096-token AIME25 trace from Qwen3-8B for all five settings. The example is taken from layer 4, KV head 5, and query head 20. The top row shows the absolute key values immediately before INT2 quantization. The bottom row shows each cached key token’s contribution to the post- W_O error, aggregated over the final 64 query positions while keeping values in BF16 to isolate key-induced error. All panels use the same sample and INT2 configuration: group size 128, key clipping ratio 0.96, 64 BF16 sink tokens, and 256 BF16 recent tokens.

Calibration details. We collect Q/K/V activation statistics from all 198 GPQA-Diamond prompts with one generated token, without using answer labels. During collection, each dynamically formed prefill batch is stored as a chunk containing aligned Q, K, V, and sequence-length tensors, yielding 14 chunks for each Qwen model and 15 for Phi-14B.

From these dumps, we use chunks 3 and 4 for optimization and chunks 10 and 12 for held-out selection. After removing sequences shorter than 328 tokens, the calibration/held-out sets contain 6/5 prompts for Qwen3-4B, 7/6 for Qwen3-8B, and 9/3 for Phi-14B, corresponding to 2,554/2,526, 2,711/3,092, and 4,151/1,619 tokens, respectively. We compute the objective over the final 64 query positions and select the best rotation for each KV head using the held-out objective every 20 steps. Key-centering statistics are estimated from 28 full model-generated traces. The calibration and held-out chunks do not overlap, although their source prompts come from the GPQA-Diamond evaluation pool.

Setting	Configuration
KV-cache precision	Group-wise affine INT2
Quantization group size	128
Key clipping ratio	0.96
Value clipping ratio	0.92
BF16 sink window	64 tokens
BF16 recent window	256 tokens
Calibration data	GPQA decoding traces
Optimizer	Adam Kingma and Ba [2014]
Optimization steps	80
Learning rate	0.02

Table A: Default experimental settings for OptR.

Experimental settings. We implement OptR on the official OSCAR codebase and adopt its best-performing INT2 configuration Zhou et al. [2026]. The default settings used throughout the main experiments and appendix are summarized in Table A. Unless explicitly stated, each ablation changes only the setting under investigation while keeping all others fixed.

We measure efficiency on a single NVIDIA A100-SXM4 40GB GPU using Qwen3-4B-Thinking-2507 with tensor parallelism of one and an eight-token page size. Standard CUDA graphs are enabled for each evaluated batch size, while piecewise CUDA graphs are disabled. Each process performs an initial untimed warm-up. For every end-to-end and prefill setting, we additionally discard one complete warm-up run before recording five timed runs. Decode latency is measured over 1,024 generated tokens and excludes prefill time. End-to-end output throughput is computed as the number of generated tokens divided by the combined prefill and decoding time. We report the mean over five runs in the main figure and omit error bars for readability. The maximum standard deviations across all evaluated settings are 0.104 ms per token for decode latency, 0.412 tokens/s for end-to-end throughput, and 0.0115 s for prefill time. We use a custom SGLang implementation, PyTorch 2.9.1 with CUDA 12.8, and Triton 3.5.1.

Key-mean estimation and application. The key mean is estimated from post-RoPE keys in the BF16 calibration traces:

$$\mu_{\ell,h} = \frac{1}{N} \sum_{n,s} k_{n,s,h}^{\ell,\text{RoPE}} \quad (19)$$

At inference, the same mean is subtracted from the post-RoPE keys before rotation and quantization. The cache-write kernel implements this operation as

$$(k_s^{\text{RoPE}} - \mu_{\ell,h}) R_{K,\ell,h} = k_s^{\text{RoPE}} R_{K,\ell,h} - \mu_{\ell,h} R_{K,\ell,h} \quad (20)$$

We precompute $\mu_{\ell,h} R_{K,\ell,h}$ once and subtract it in the rotated space. BF16 sink and recent keys retain the same centering.

Rotated-query implementation. For clarity, the formulation in the main paper maps an effective quantized key back with $R_K^\top$. Let

$$k_{R,s} = Q_2((k_s - \mu) R_K) \quad (21)$$

The corresponding effective key is $\tilde{k}_s = k_{R,s} R_K^\top$. Its attention logit satisfies

$$q_t \tilde{k}_s^\top = q_t R_K k_{R,s}^\top \quad (22)$$

Therefore, the implementation stores $k_{R,s}$ directly in the INT2 cache and applies R_K to the post-RoPE query instead of explicitly applying $R_K^\top$ to every dequantized key. Under GQA, each query head uses the rotation of its corresponding KV head. The two implementations produce identical attention logits.

Equivalent Inference Implementation. Figure 3 shows the baseline and output-aware correction rotations as separate operations. We denote their combined key and value rotations by

$$R_K = R_k R_{kc} \quad R_V = R_v R_{vc} \quad (23)$$

For the key path, the same R_K is applied to the post-RoPE query and centered key. Since R_K is orthogonal, the full-precision attention logit satisfies

$$(q_t R_K) ((k_s - \mu) R_K)^\top = q_t (k_s - \mu)^\top \quad (24)$$

Thus, the key rotations cancel in the full-precision inner product. The remaining term $-q_t \mu^\top$ is constant across all cached tokens s and therefore leaves the softmax distribution unchanged.

This equivalence holds before quantization. Under INT2, the stored key is

$$k_{R,s} = Q_2((k_s - \mu) R_K) \quad (25)$$

and the attention logit becomes $(q_t R_K) k_{R,s}^\top$. Therefore, the rotation changes the coordinate system in which the INT2 error is introduced, although it does not change the full-precision attention computation. This implementation is equivalent to the inverse-rotation formulation in the main paper:

$$q_t (k_{R,s} R_K^\top)^\top = (q_t R_K) k_{R,s}^\top \quad (26)$$

We therefore store $k_{R,s}$ directly in the INT2 cache and apply R_K to each query head using the rotation of its corresponding KV head.

For the value path, Figure 3 explicitly applies R_v and R_{vc} before quantization and their inverses after attention. With $R_V = R_v R_{vc}$, the rotated value and attention output are

$$v'_s = v_s R_V \quad o'_j = o_j R_V \quad (27)$$

The explicit inverse path recovers the original output:

$$o'_j R_{vc}^{-1} R_v^{-1} W_{O,j}^\top = o_j W_{O,j}^\top \quad (28)$$

In the implementation, these value-side transforms are folded into the projection weights. For KV head h and query head $j \in G_h$, we use

$$\overline{W}_{V,h} = R_{V,h}^\top W_{V,h} \quad \overline{W}_{O,j} = W_{O,j} R_{V,h} \quad (29)$$

These weights satisfy

$$x \overline{W}_{V,h}^\top = (x W_{V,h}^\top) R_{V,h}, \quad (30)$$

$$(o_j R_{V,h}) \overline{W}_{O,j}^\top = o_j W_{O,j}^\top \quad (31)$$

Hence, the value rotations and their inverses shown in Figure 3 are mathematical operations rather than separate inference kernels.

Models and Benchmarks. We evaluate OptR across three reasoning-oriented language models that differ in model scale and architecture. This setting allows us to examine whether the proposed method remains effective across both compact and larger models, as well as across distinct model families.

- **Qwen3-4B-Thinking-2507 and Qwen3-8B** Yang et al. [2025]: two Qwen3 models at different scales, used to assess whether the effectiveness of OptR is preserved as model scale increases.
- **Phi4-14B-reasoning-plus** Abdin et al. [2025]: a reasoning model from a different model family, included to evaluate cross-architecture generalization.

Together, these models cover parameter scales from 4B to 14B and include both within-family scaling and cross-family evaluation.

We evaluate reasoning, coding, and long-context retrieval capabilities to measure the effect of INT2 KV-cache quantization across diverse inference workloads.

- **AIME24 and AIME25** MAA [2025]: challenging mathematical reasoning benchmarks that require multi-step problem solving.

- **GPQA-Diamond** Rein et al. [2024]: a graduate-level scientific reasoning benchmark covering questions that require specialized knowledge and careful reasoning.
- **MBPP+** Liu et al. [2023]: a code-generation benchmark with extended test cases for more rigorous functional-correctness evaluation.
- **LiveCodeBench v6** Jain et al. [2025]: a contamination-resistant coding benchmark constructed from recent programming problems.
- **RULER-NIAH** Hsieh et al. [2024]: a long-context retrieval benchmark used to evaluate whether quantized KV caches preserve information over extended context lengths.

The reasoning and coding benchmarks evaluate the quality of generated outputs, whereas RULER-NIAH isolates long-context retrieval performance under increasing context lengths.

Generation Settings. We use the same sampling configuration across all models and benchmarks to ensure a consistent comparison. Specifically, we set the temperature to 0.6, top- p to 0.95, and top- k to 20. For the reasoning and coding benchmarks, we use maximum generation lengths of 32K tokens for the Qwen3 models and 16K tokens for Phi4-14B-reasoning-plus.

B Output-Error Measurement and Analysis

This section describes the measurement settings for the per-layer line plots, layer-averaged bar plots, and per-head heatmaps. The main paper reports the Qwen3-4B-Thinking-2507 per-layer analysis and GPQA-Diamond heatmap in Figures 4 and 5, respectively. Additional per-layer and layer-averaged results for Qwen3-4B-Thinking-2507 and Qwen3-8B are provided in Figures B–E, while the complete per-head heatmaps across both models and datasets are shown in Figure A. Unless otherwise specified, all analyses use the same INT2 quantization settings as the main experiments.

Output-space error. For each layer ℓ , KV head h , and query head $j \in \mathcal{G}_h$, we use the last 64 query positions of each sequence. Let $p_t^{\ell,j,h}$ and $\tilde{p}_t^{\ell,j,h}$ denote the causal attention distributions obtained with full-precision and INT2 keys, respectively. We write $V_{1:t,h}^\ell$ and $\tilde{V}_{1:t,h}^\ell$ for the full-precision and effective INT2 value matrices, where the latter includes full-precision sink- and recent-token restoration. Following the notation in the main paper, the post- W_O output discrepancy is

$$\Delta y_{t,j}^\ell = \left[(\tilde{p}_t^{\ell,j,h})^\top \tilde{V}_{1:t,h}^\ell - (p_t^{\ell,j,h})^\top V_{1:t,h}^\ell \right] (W_{O,j}^\ell)^\top$$

This quantity is the total output error $\Delta y_{t,j}^\ell = \tilde{y}_{t,j}^\ell - y_{t,j}^\ell$ and jointly captures key-induced changes in the attention distribution and value-induced changes in the attention-weighted output. For each layer, we average $\|\Delta y_{t,j}^\ell\|_2^2$ over query positions, query heads, KV heads, and sequences. We apply the same group-wise INT2 quantization and dequantization path used during inference to the KV cache of the measured layer while using full-precision input activations. The resulting values measure the immediate error introduced by each attention block. The corresponding per-layer results are reported in Figures 4, B, and D.

Propagated residual-stream error. To measure the accumulation of quantization error across depth, we run the complete model with INT2 KV caches in all Transformer layers. For each of 12 prompts per dataset, the full-precision model first generates a reasoning trace greedily, with up to 1536 new tokens and a maximum sequence length of 3072. The full-precision and INT2 models are then evaluated on the same generated token sequence. This teacher-forced comparison isolates representational drift from differences caused by autoregressive sampling.

All model weights remain in BF16. Long-history keys and values are quantized and stored in INT2 at cache-write time and dequantized during attention, while the configured sink- and recent-token windows remain in full precision. Let $h_{\ell,t}$ and $\tilde{h}_{\ell,t}$ denote the full-precision and INT2-KV residual streams after block ℓ at token position t . We report the relative squared error

$$\frac{\sum_t \|\tilde{h}_{\ell,t} - h_{\ell,t}\|_2^2}{\sum_t \|h_{\ell,t}\|_2^2}$$

pooled over all tokens and prompts. This normalization accounts for changes in the residual-stream magnitude across depth and enables comparison across layers. The propagated errors across model depth are shown in panels (c,d) of Figures 4, B, and D.

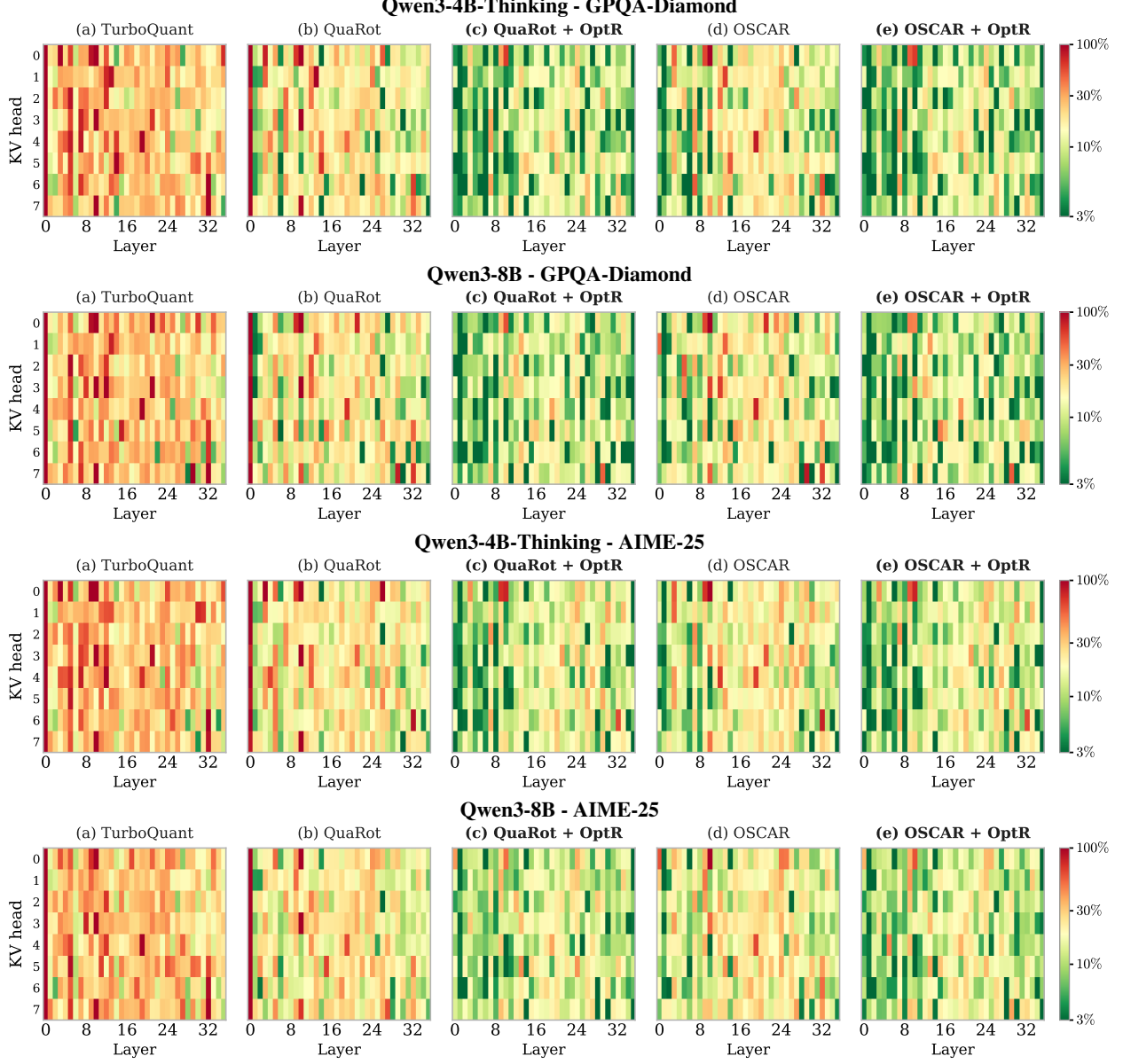


Figure A: Output error by layer and KV head relative to naive INT2 across models and evaluation datasets. Each cell reports the post- W_O attention-output RMS error as a percentage of plain per-group INT2 without rotation. Lower values indicate smaller output error, with naive INT2 corresponding to 100%. Within each heatmap, panels show (a) TurboQuant, (b) QuaRot, (c) QuaRot + OptR, (d) OSCAR, and (e) OSCAR + OptR.

Layer-averaged summaries. The bar plots summarize the corresponding per-layer line plots by reporting the arithmetic mean of each error metric across Transformer layers. Panels (a,b) report the mean post- W_O output-space MSE on GPQA-Diamond and AIME25, while panels (c,d) report the corresponding mean propagated residual-stream error. The Qwen3-4B-Thinking-2507 and Qwen3-8B summaries are reported in Figures C and E, respectively. These plots provide an aggregate comparison between methods, whereas Figures B and D show how the same errors vary across model depth.

Per-head error relative to naive INT2. The heatmaps use the same post- W_O output error $\Delta y_{t,j}^\ell$, but retain a separate value for each (layer, KV head) pair rather than averaging over KV heads. Let $\text{err}_{\ell,h}^{\text{method}}$ denote the mean squared output error for a given cell and let $\text{err}_{\ell,h}^{\text{INT2}}$ denote the corresponding error under plain per-group INT2 without rotation

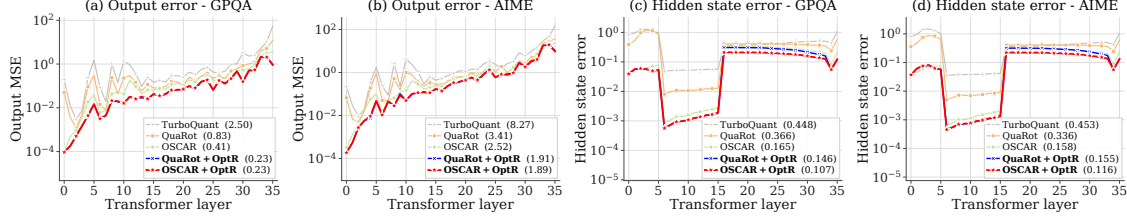


Figure B: Per-layer error analysis under INT2 KV-cache quantization on Qwen3-4B-Thinking. Panels (a,b) report the post- W_O output-space MSE on GPQA-Diamond and AIME25, capturing the immediate error introduced by each attention block. Panels (c,d) report the propagated residual-stream error, computed as the normalized squared difference between full-precision and INT2-KV block-output hidden states evaluated on the same full-precision-generated sequences. Values in the legend denote averages across layers. All errors are shown on a logarithmic scale.

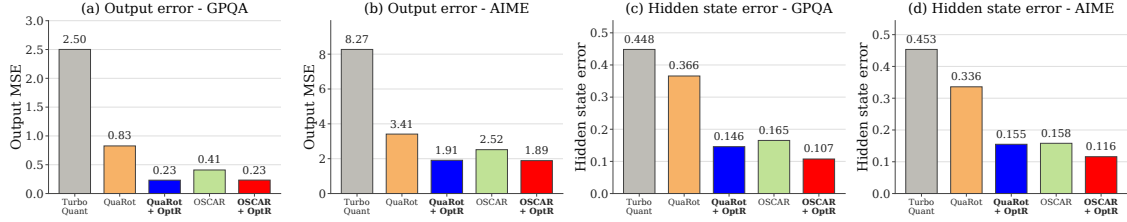


Figure C: Layer-averaged error under INT2 KV-cache quantization on Qwen3-4B-Thinking. Panels (a,b) report the mean post- W_O output-space MSE across Transformer layers on GPQA-Diamond and AIME25, respectively. Panels (c,d) report the corresponding mean propagated residual-stream error. The values summarize the per-layer results shown in Figure B. Lower values indicate smaller quantization-induced error.

or clipping. Each cell reports $100\sqrt{\text{err}_{\ell,h}^{\text{method}}/\text{err}_{\ell,h}^{\text{INT2}}}$, representing the RMS output error as a percentage of naive INT2. Thus, naive INT2 corresponds to 100%, and lower values indicate smaller output error. The heatmaps use a logarithmic scale ranging from 3% to 100%. The GPQA-Diamond and AIME25 heatmaps are shown in Figures A.

Results. Figures B and C show that OptR reduces both immediate post- W_O output error and propagated residual-stream error on Qwen3-4B-Thinking-2507. On GPQA-Diamond, OptR reduces the average post- W_O output MSE from 0.83 to 0.23 for QuaRot and from 0.41 to 0.23 for OSCAR. On AIME25, the corresponding errors decrease from 3.41 to 1.91 and from 2.52 to 1.89, respectively. The propagated residual-stream error decreases from 0.366 to 0.146 for QuaRot and from 0.165 to 0.107 for OSCAR on GPQA-Diamond, and from 0.336 to 0.155 and from 0.158 to 0.116 on AIME25.

The same trend holds for Qwen3-8B, as shown in Figures D and E. On GPQA-Diamond, OptR reduces the average post- W_O output MSE from 2.31 to 0.90 for QuaRot and from 1.73 to 0.87 for OSCAR. On AIME25, the corresponding errors decrease from 8.88 to 4.00 and from 7.02 to 4.29. The propagated residual-stream error decreases from 0.195 to 0.127 for QuaRot and from 0.147 to 0.109 for OSCAR on GPQA-Diamond. On AIME25, it decreases from 0.210 to 0.131 for QuaRot and from 0.135 to 0.132 for OSCAR.

Finally, Figure A shows that the error reductions extend across most layers and KV heads on both GPQA-Diamond and AIME25. The improvements are therefore broadly distributed throughout the models rather than being concentrated in a small subset of layers or heads.

Tokens	OSCAR	OSCAR + OptR
3.1k	62.00 $\pm$ 6.50	69.33 $\pm$ 3.65
6.6k	63.33 $\pm$ 4.08	70.67 $\pm$ 2.79
12.6k	62.67 $\pm$ 5.96	70.00 $\pm$ 5.77
19.2k	64.00 $\pm$ 3.65	70.67 $\pm$ 4.35
26.0k	62.67 $\pm$ 6.41	70.00 $\pm$ 4.08

Table B: Effect of calibration size on AIME25 accuracy for Qwen3-4B-Thinking-2507. Both methods use the same GPQA calibration tokens and a BF16 window of $S=64/R=256$. Results are $\mu \pm \sigma$ over 5 seeds.

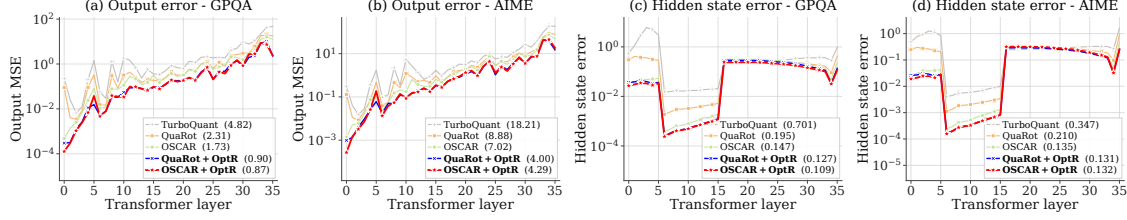


Figure D: Per-layer error analysis under INT2 KV-cache quantization on Qwen3-8B. Panels (a,b) report the post- W_O output-space MSE on GPQA-Diamond and AIME25, capturing the immediate error introduced by each attention block. Panels (c,d) report the propagated residual-stream error, computed as the normalized squared difference between full-precision and INT2-KV block-output hidden states evaluated on the same full-precision-generated sequences. Values in the legend denote averages across layers. All errors are shown on a logarithmic scale.

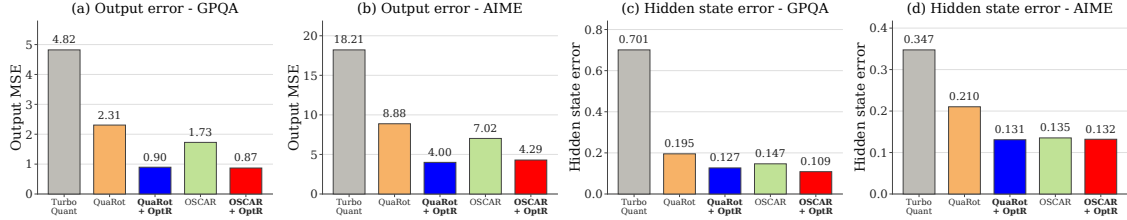


Figure E: Layer-averaged error under INT2 KV-cache quantization on Qwen3-8B. Panels (a,b) report the mean post- W_O output-space MSE across Transformer layers on GPQA-Diamond and AIME25, respectively. Panels (c,d) report the corresponding mean propagated residual-stream error. The values summarize the per-layer results shown in Figure D. Lower values indicate smaller quantization-induced error.

B.1 Ablation on Calibration Size

Table B evaluates the sensitivity of OSCAR and OptR to the number of calibration tokens. OptR consistently improves OSCAR across all calibration sizes, while its performance remains stable from a few thousand tokens onward. Increasing the calibration size beyond 6.6k tokens provides no consistent additional gain, indicating that OptR requires only a modest calibration set. We therefore use 6.6k tokens as the default configuration.

B.2 Ablation on λ_K

Table C studies the sensitivity of OptR to the key-loss weight λ_K . The best value depends on the base rotation: OSCAR + OptR performs best at $\lambda_K = 1$, whereas QuaRot-INT2 + OptR achieves its highest accuracy at $\lambda_K = 10$.

λ_K	QuaRot-INT2 + OptR	OSCAR + OptR
0.1	60.67 $\pm$ 5.48	66.67 $\pm$ 4.08
0.3	62.00 $\pm$ 4.35	63.33 $\pm$ 5.27
1	60.67 $\pm$ 4.35	70.67 $\pm$ 2.79
3	60.67 $\pm$ 4.35	63.33 $\pm$ 4.71
10	63.33 $\pm$ 4.08	60.67 $\pm$ 2.79

Table C: Effect of λ_K on AIME25 accuracy for Qwen3-4B-Thinking-2507. Results are $\mu \pm \sigma$ over five seeds, and the shaded row denotes the default setting.

This difference indicates that the appropriate balance between attention-distribution preservation and post- W_O output-error reduction depends on the rotation initialization. To avoid base-specific tuning, we use $\lambda_K = 1$ throughout the main experiments, which gives the strongest performance with OSCAR, our state-of-the-art rotation baseline.

Offline calibration cost. Table D reports the one-time offline calibration cost of OptR. Calibration requires only several minutes on a single GPU across the evaluated model sizes. All model weights remain frozen, and the resulting rotations are reused across subsequent inference requests, amortizing this cost during deployment.

Model	Calibration Time (Minute)
Qwen3-4B-Thinking-2507	6.25
Qwen3-8B	7.05
Phi4-14B-Reasoning-Plus	11.03

Table D: Calibration time of OptR with 80 optimization steps. Measurements use a single NVIDIA A100 GPU and exclude BF16 trace collection. All model weights remain frozen throughout calibration.

Model	Method	4k	8k	16k	32k	64k	128k
Qwen3-8B	BF16	99.83 $\pm$ 0.11	99.93 $\pm$ 0.02	99.45 $\pm$ 0.08	98.70 $\pm$ 0.49	84.22 $\pm$ 1.41	80.89 $\pm$ 0.58
	QuaRot-INT2	84.97 $\pm$ 0.09	41.73 $\pm$ 5.75	18.16 $\pm$ 1.14	13.03 $\pm$ 1.71	0.04 $\pm$ 0.07	0.38 $\pm$ 0.13
	QuaRot-INT2 + OptR	99.40 $\pm$ 0.29	98.76 $\pm$ 0.57	96.14 $\pm$ 0.35	86.37 $\pm$ 0.72	70.02 $\pm$ 1.79	41.87 $\pm$ 0.96
	OSCAR	99.59 $\pm$ 0.15	97.94 $\pm$ 0.28	94.39 $\pm$ 0.29	83.76 $\pm$ 0.58	57.54 $\pm$ 1.82	26.44 $\pm$ 0.87
	OSCAR + OptR	99.60 $\pm$ 0.11	98.16 $\pm$ 0.63	95.50 $\pm$ 0.65	86.42 $\pm$ 0.55	68.65 $\pm$ 0.80	40.43 $\pm$ 1.53

Table E: RULER-NIAH retrieval accuracy of Qwen3-8B at context lengths up to 128K tokens. Results are $\mu \pm \sigma$ over three seeds, with 800 examples per seed. Bold denotes the best INT2 result at each context length.

Long-Context Results Table E extends the Qwen3-8B long-context evaluation to 128K tokens. At this context length, QuaRot-INT2 and OSCAR degrade substantially, whereas applying OptR retains considerably higher retrieval accuracy. The consistent gains with both base rotations show that OptR remains effective beyond the 64K range reported in the main paper.