QK-WANDA: COUPLING QUERIES AND KEYS FOR UNSTRUCTURED PRUNING

Anonymous authors
Paper under double-blind review

ABSTRACT

Pruning aims to reduce the storage and computational costs of large language models by removing selected weights. WANDA (Sun et al., 2024) scores query and key weights independently, although their outputs interact through dot products in attention. We introduce QK-WANDA, which scores each weight by the exact reconstruction loss caused by deleting it alone. The objective compares query–key products before rotary position embeddings (RoPE), retaining only causally allowed token pairs. The resulting closed-form scores use activations from the opposite projection and support a shared Q/K budget without gradients or updates to retained weights. We evaluate Q/K-only pruning on 15 checkpoints from TinyLlama, Llama 2, Llama 3, and Qwen2.5, spanning 0.5B–72B parameters. Relative to row-wise Wanda, shared-budget QK-Wanda reduces held-out relative squared QK error by approximately 60% at 50% Q/K sparsity and 45% at 80%, averaging relative reductions across checkpoints. Downstream benefits are model-dependent. At 80% Q/K sparsity on Llama 2 70B, WikiText-2 and C4 perplexity decrease by 20.2% and 13.5%, while mean zero-shot accuracy increases by 5.91 percentage points. Qwen2.5-72B also improves at this sparsity, whereas Llama-3.1-70B has substantially higher perplexity despite lower reconstruction error. These results demonstrate both the potential of coupled pruning criteria and the limits of local reconstruction as a predictor of downstream quality.

1 INTRODUCTION

Large language models (LLMs) perform well across language understanding, generation, and reasoning tasks, but storing and running their billions of parameters remains expensive (Touvron et al., 2023; Grattafiori et al., 2024; Qwen et al., 2024). Model compression addresses this cost. Quantization reduces the bits per weight; pruning sets selected weights to zero. Both seek to preserve predictive quality. We focus on unstructured pruning (Hoefler et al., 2021).

WANDA selects weights using their magnitudes and calibration activations, without updating retained weights (Sun et al., 2024). We focus on this mask-only setting, rather than methods that also compensate for pruning errors, such as SPARSEGPT and THANOS (Frantar & Alistarh, 2023; Ilin & Richtárik, 2025). Our question is whether accounting for query–key interactions can improve these scalar pruning decisions.

Let $W \in \mathbb{R}^{m\times d}$ be a weight matrix and $X \in \mathbb{R}^{d\times T}$ its calibration inputs, with one token per column. The entry W_{ij} connects input feature j to output coordinate i; the colon in $X_{j:}$ selects feature j across all T tokens. We write Wanda's score in squared form:

$$S_{ij}^{W} = W_{ij}^{2}\,\|X_{j:}\|_2^2, \tag{1}$$

where $\|X_{j:}\|_2$ is the Euclidean norm and the superscript W labels Wanda. Squaring its original nonnegative score preserves the ranking. Smaller scores are pruned first, with a fixed removal fraction in each output row. This score equals the squared change in the linear output WX caused by deleting one weight while holding the others fixed.

Wanda measures each deletion through its effect on an individual projection. In attention, however, a query perturbation changes dot products according to the keys it interacts with, and vice versa (Vaswani et al., 2017). QK-WANDA therefore scores deletions using a common reconstruction

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.

objective for causally allowed query–key products before rotary position embeddings. The resulting query and key scores measure changes to the same objective, allowing a fixed pruning budget to be allocated across rows and between the two projections.

We first evaluate Q/K-only pruning, then combine QK-Wanda with Wanda on the remaining transformer-block projections.

Contributions.

1. **Causal scalar-deletion scores.** We derive closed-form costs for deleting individual query or key weights under a causal pre-RoPE reconstruction objective (Equation (7)), including shared-key heads and retained biases (Equations (12) and (13)). We also derive a RoPE-aware variant.
2. **Coupled mask selection without weight updates.** We compute the scores using prefix/suffix accumulators and select masks under separate or shared Q/K budgets, without gradients or changes to retained weights.
3. **Controlled evaluation of scoring and allocation.** Across 15 checkpoints, we evaluate reconstruction, perplexity, and zero-shot accuracy. Matched-allocation controls show that gains are not solely due to relaxing Wanda's row quotas. The results demonstrate substantial high-sparsity improvements on TinyLlama and Llama 2, while identifying regimes where better local reconstruction does not preserve downstream quality.

2 Background and Related Work

Weight importance and reconstruction. Optimal Brain Damage uses diagonal second-order loss sensitivity; Optimal Brain Surgeon includes compensating weight changes (LeCun et al., 1989; Hassibi & Stork, 1992). Optimal Brain Compression develops layer-reconstruction methods for pruning and quantization (Frantar et al., 2022). SparseGPT scales compensated pruning to LLMs, while Thanos solves for several deletions within a weight block together and adapts their allocation (Frantar & Alistarh, 2023; Ilin & Richtárik, 2025). We study mask selection with surviving weights fixed.

Activation-aware scalar scores. Wanda's magnitude–activation score measures a deletion's effect on one linear projection, as expressed in Equation (1) (Sun et al., 2024). SymWanda extends this view with input- and output-side reconstruction terms and derives scalar scores for a family of objectives (Yi & Richtárik, 2025). These input- and output-side terms reconstruct a single matrix. Our objective couples query and key projection weights through dot products between their output activations on causally allowed token pairs. The resulting deletion costs are comparable across Q/K.

Attention-aware quantization. AESPA and BoA use coupled Q/K reconstruction for quantization, with BoA additionally accounting for RoPE (Kim et al., 2024; 2025). We study scalar pruning instead: each score measures the finite change caused by deleting an original weight, retaining causal token-pair restrictions and leaving the other weights unchanged. Appendix K details the relationship between the reconstruction objectives.

Structured Q/K and attention compression. CORP ranks shared Q/K dimensions by their logit contribution energy and applies closed-form compensation (Zhang & Yang, 2026). This removes matched Q/K output rows, whereas our scalar deletion leaves every other coordinate fixed. SNP likewise selects paired Q/K neurons, using alignment with singular components of attention scores (Shim et al., 2024). SparseLoRA uses a QK norm criterion for dynamic channel selection during fine-tuning (Khaki et al., 2025). Their pruning units and recovery or training procedures differ from ours.

Other approaches change the representation or reconstruction target. Liang et al. (2025) optimize a mask on the fused QK weight map through a nonlinear attention objective. CLOVER prunes orthogonal directions, A^3 derives low-rank attention approximations, and RAP removes RoPE-aligned channel pairs for KV-cache compression (Meng et al., 2025; Wong et al., 2025; Xin et al., 2026). LLM-Pruner, SliceGPT, and FASP provide broader structural alternatives; DIMAP develops a

data-independent module-aware criterion for vision transformers (Ma et al., 2023; Ashkboos et al., 2024; Hu et al., 2025; He & Zhou, 2024). These decisions differ from our scalar mask on the original projection matrices.

Allocation, calibration, and attention semantics. OWL studies nonuniform layerwise sparsity (Yin et al., 2024), while Ji et al. (2025) examine sensitivity to calibration data. We study budget allocation between Q and K within a block and vary calibration size and sequence length. Causal masking, shared key heads, and rotary positions affect which Q/K interactions a score measures (Vaswani et al., 2017; Ainslie et al., 2023; Su et al., 2024).

Other quantization approaches reconstruct individual layers or optimize compressed representations, including GPTQ, AQLM, PV-TUNING, HIGGS, and QTIP (Frantar et al., 2023; Egiazarian et al., 2024; Malinovskii et al., 2024a;b; Tseng et al., 2024).

3 QK-WANDA

3.1 THE CORE IDEA: AN UNMASKED QK RECONSTRUCTION LOSS

We first consider single-head attention and one calibration sequence. This unmasked, pre-RoPE case, QK-WANDA-U, makes the coupling especially simple: a query weight matters through the keys, and a key weight through the queries. We then add the causal restriction to obtain our main method, QK-WANDA.

Let $X \in \mathbb{R}^{d\times T}$ contain one input token per column, and let $W^Q, W^K \in \mathbb{R}^{m\times d}$ be the query/key projection weights. Biases $b^Q, b^K \in \mathbb{R}^m$ remain fixed and are zero for bias-free models. With the all-ones column vector $e \in \{1\}^{T\times 1}$, the pre-RoPE activations are

$$Q = W^Q X + b^Q e^\top, \qquad K = W^K X + b^K e^\top. \tag{2}$$

Here $Q, K \in \mathbb{R}^{m\times T}$; the row $Q_{i:}$ or $K_{i:}$ contains coordinate i across all tokens.

We compare the query–key products directly. Hats on Q and K denote their values after pruning, and the squared Frobenius norm $\|M\|_F^2$ is the sum of the squared entries of a matrix M. For the unmasked, pre-RoPE case, define

$$\mathcal{L} = \left\| \widehat{Q}^\top \widehat{K} - Q^\top K \right\|_F^2. \tag{3}$$

Every token pair contributes, and inputs are fixed during a pruning decision. Attention divides its dot products by $\sqrt{m}$ before softmax. We omit this scale from the pruning losses and scores derived below: including it would divide every deletion cost by m, leaving the ranking unchanged. The model's forward pass retains its original scaling.

Deleting one query weight. Set W^Q_{ij} to zero, keeping keys, biases, and every other weight fixed. Write $\Delta Q = \widehat{Q} - Q$ for the resulting change. The query bias cancels in this difference; only query row i changes, by $\Delta Q_{i:} = -W^Q_{ij} X_{j:}$. Therefore

$$\widehat{Q}^\top K - Q^\top K = \Delta Q^\top K = -W^Q_{ij} X_{j:}^\top K_{i:}. \tag{4}$$

This is an outer product, so $\left\|xy^\top\right\|_F^2 = \|x\|_2^2 \, \|y\|_2^2$ gives its squared reconstruction cost. By symmetry, deleting W^K_{ij} with queries fixed yields $Q^\top \Delta K = -W^K_{ij} Q_{i:}^\top X_{j:}$. The two exact single-deletion scores are

$$\boxed{S^Q_{ij} = \underbrace{(W^Q_{ij})^2 \, \|X_{j:}\|_2^2}_{\text{WANDA}} \|K_{i:}\|_2^2, \qquad S^K_{ij} = \underbrace{(W^K_{ij})^2 \, \|X_{j:}\|_2^2}_{\text{WANDA}} \|Q_{i:}\|_2^2.} \tag{5}$$

Each score multiplies the squared Wanda score by the opposite projection's coordinate energy; smaller values indicate less reconstruction damage. These are finite-deletion identities, without a small-change approximation. A bias cancels from its own perturbation but remains in the opposite activation: the key bias affects the query score through K, and the query bias affects the key score through Q.

3.2 ADDING THE CAUSAL MASK: QK-WANDA

The unmasked score captures Q/K coupling, but also penalizes changes to future-token pairs that a decoder never uses. We refine the objective by retaining only allowed pairs: query position t can attend to key position u exactly when $u \le t$. Both positions range from 1 to T. We call this causal, pre-RoPE version QK-WANDA, our main method.

For the same one-sequence, one-head setting, define

$$\mathcal{L}_C = \left\| C \odot (\widehat{Q}^\top \widehat{K} - Q^\top K) \right\|_F^2, \qquad C_{tu} = \mathbf{1}\{u \le t\}, \tag{6}$$

where $\odot$ denotes entrywise multiplication and the indicator $\mathbf{1}\{u \le t\}$ is one for allowed pairs and zero otherwise. The score is the loss caused by one deletion. Write $K_{i,1:t}$ for key coordinate i at positions $1, \dots, t$, and $Q_{i,u:T}$ for query coordinate i at positions $u, \dots, T$. Let $\hat{S}^Q_{ij}$ and $\hat{S}^K_{ij}$ denote the causal costs of deleting W^Q_{ij} and W^K_{ij}, respectively. A hat on S denotes the causal score. They are exactly

$$\boxed{\hat{S}^Q_{ij} = (W^Q_{ij})^2 \sum_{t=1}^{T} X^2_{jt} \left\| K_{i,1:t} \right\|_2^2, \qquad \hat{S}^K_{ij} = (W^K_{ij})^2 \sum_{u=1}^{T} X^2_{ju} \left\| Q_{i,u:T} \right\|_2^2.} \tag{7}$$

The query score uses *key-prefix energy*: the squared norm of keys available to that query token. The key score uses *query-suffix energy*: the squared norm of all query tokens that can attend to that key. Unlike the unmasked case, the opposite-side factor depends on token position and cannot generally be taken outside the sum.

With grouped-query attention (GQA), deleting a key weight changes every query head that shares it, so its cost sums those heads' contributions. We do not divide key scores by the number of associated query heads: Q and K scores must retain the same summed-head loss scale for shared allocation. Proofs, multiple-sequence formulas, and a two-token example are in Sections A.1 and A.2; Section B.1 gives the unmasked extension.

Reciprocal rescaling of a query group and its shared key projection leaves their products and QK-Wanda scores unchanged, provided the associated biases are scaled as well (Appendix A).

QK-WANDA-R applies the same scalar-deletion construction to a causal target that includes RoPE. Section 4.6 compares the variants and motivates the pre-RoPE default; Appendix B.2 gives the derivation. All variants retain the model's original forward computation.

For multiple deletions, cross terms appear within each projection and between Q and K. Thus the scores couple Q and K, but scalar thresholding does not jointly optimize both masks. Section A gives the quadratic form and the interactions omitted by summing individual scores.

3.3 MASK CONSTRUCTION AND COMPUTATIONAL COST

Why relax row-wise budgets? The score estimates the cost of a deletion, whereas the budget determines which weights can compete for removal. Fixed row quotas prevent deletions from moving between rows with different reconstruction costs. In the single-sequence unmasked setting of Equation (5), this restriction is particularly consequential: a positive opposite-projection factor only rescales Wanda's scores within each row, leaving the mask unchanged under the same tie rule. We therefore allow competition across rows and, with a shared budget, across Q and K. Causal weighting and multiple calibration sequences can also change within-row rankings (Appendix A.3).

Let $\rho \in [0, 1]$ be the requested fraction of weights to remove. We prune transformer blocks sequentially. For each block, we collect activations before modifying its Q/K weights, compute both score arrays, and construct both masks. With separate budgets, we remove the lowest-scoring fraction ρ from each projection; with a shared budget, we remove that fraction from their combined weights. We then apply both masks and propagate the sparse block output to the next block. Budgets are block-local and impose neither row nor head quotas. Integer counts and tie rules are specified in Appendix D.5.

Here "one-shot" denotes one scoring and masking decision per projection; later blocks see the effects of previous pruning. For matched calibration settings, the paired methods use identical sampled input tokens, while their later hidden states can differ.

Prefix/suffix scans are running sums from the start/end of a sequence. Along with matrix products, they compute all QK-WANDA scores without constructing a separate $T \times T$ error matrix for each weight. In the single-head setting above, the scans cost $O(T(d+m))$ and the interaction products cost $O(Tmd)$, followed by elementwise score formation and sorting. The method requires no gradients, inner optimization, or survivor updates; Section L gives the general computational cost and runtime measurements.

4 EXPERIMENTS

4.1 EXPERIMENTAL SETUP

Models and pruning. We evaluated 15 base checkpoints: TinyLlama-1.1B; Llama 2 7B/13B/70B; Llama-3.2-1B/3B and Llama-3.1-8B/70B; and Qwen2.5 0.5B/1.5B/3B/7B/14B/32B/72B (Zhang et al., 2024; Touvron et al., 2023; Grattafiori et al., 2024; Meta AI, 2024; Qwen et al., 2024). Llama 2 7B/13B use multi-head attention (MHA); the other checkpoints use grouped-query attention (GQA). TinyLlama is independently trained, not an official small Llama 2 checkpoint. Unless stated otherwise, sparsity refers only to W^Q, W^K; all other parameters remain dense. Q/K weights comprise 4.46–16.11% of total model parameters across these checkpoints, so 50% Q/K sparsity removes 2.23–8.06% of the total (Appendix D.4). We compared original row-wise WANDA with causal pre-RoPE QK-WANDA, using shared Q/K budgets per block in tables and also separate matrix budgets in curves. On the three largest checkpoints, we also tested Wanda with a separate matrix-wide budget for each projection. All methods used sequential, one-shot unstructured pruning without retraining or survivor updates; biases and the original attention computation remained fixed. Integer budgets and reproducibility details are in Appendix D.

Calibration. The main comparisons used calibration seed 0 and 524,288 C4 training tokens (Raffel et al., 2020): 256 windows of length 2048, segmented into 8192 independent sequences of length 64. Calibration settings and the shared-budget default were selected using the reported WikiText-2 test and C4 validation PPL on Llama-3.1-8B, without a separate development split. We then held this configuration fixed across the remaining checkpoints, evaluating transfer across models on the same benchmarks. Each checkpoint used its native tokenizer and identical sampled tokens across methods and sparsities. We used FP16 weights except for Qwen2.5-72B, whose dense and pruned models used BF16 to avoid nonfinite FP16 activations. Hardware and software details are in Appendix D.5.

Metrics and comparisons. We evaluate pruning at two levels. Held-out QK reconstruction error measures how well each block preserves the local objective on common dense inputs. Perplexity on WikiText-2 (Merity et al., 2017) and C4 and mean accuracy across seven zero-shot tasks measure the quality of the complete pruned model. Comparing these metrics tests whether better local reconstruction translates into downstream gains.

QK error is the mean across blocks of relative squared causal pre-RoPE reconstruction error, using 16 held-out C4 sequences of length 512 (Equation (47)). This diagnostic differs from pruning calibration, where later blocks receive method-specific sparse-network inputs. We measured WikiText-2 test and C4 validation PPL at context 2048 over 0–90% sparsity in 10-point steps. At 30%, 50%, and 80%, we also evaluated the saved row-Wanda/shared-QK masks and the large-model matrix-Wanda controls on seven zero-shot tasks, reporting their unweighted mean accuracy with LM Evaluation Harness (EleutherAI, 2026); tasks, splits, and metrics are specified in Appendix D.2. Dense references are shared within each checkpoint. The 15-model counts compare shared QK-Wanda against original row-wise Wanda, using unrounded values and excluding dense points: 135 checkpoint–sparsity pairs per PPL/error metric and 45 for mean accuracy. They describe one calibration seed, not independent statistical trials.

4.2 LARGEST CHECKPOINTS ACROSS MODEL FAMILIES

Table 1: Q/K-only pruning of the largest checkpoints: WikiText-2/C4 PPL (↓), seven-task mean accuracy (↑), and relative squared causal pre-RoPE QK error (↓). Accuracy and error are percentages. Wanda uses row budgets; Wanda (matrix) uses separate projection budgets; QK-Wanda uses shared Q/K budgets. Bold marks the best displayed sparse results, including displayed ties. Full family tables and task scores are in Appendices F, G, and M.

		Llama-2-70B				**Llama-3.1-70B**				**Qwen2.5-72B**			
Method	Sparsity	Wiki2 PPL	C4 PPL	Zero-shot acc.	Rel. QK error	Wiki2 PPL	C4 PPL	Zero-shot acc.	Rel. QK error	Wiki2 PPL	C4 PPL	Zero-shot acc.	Rel. QK error
Dense	0%	3.32	5.71	71.50	0.00	2.81	7.11	74.21	0.00	3.87	9.26	74.75	0.00
WANDA	30%	**3.32**	**5.71**	71.50	0.04	**2.82**	**7.12**	74.16	0.04	**3.88**	**9.26**	74.68	0.03
WANDA (matrix)	30%	**3.32**	**5.71**	71.49	0.05	2.86	7.15	73.95	0.17	3.89	**9.26**	**75.02**	0.05
QK-WANDA	30%	**3.32**	**5.71**	**71.62**	**0.01**	**2.82**	**7.12**	**74.18**	**0.01**	**3.88**	**9.26**	74.61	**0.01**
WANDA	50%	3.36	5.74	71.37	0.46	**2.90**	**7.22**	73.60	0.37	3.92	9.28	**74.72**	0.18
WANDA (matrix)	50%	3.36	5.74	**71.80**	0.56	3.35	7.61	72.54	1.58	3.92	9.29	74.59	0.28
QK-WANDA	50%	**3.33**	**5.72**	71.62	**0.06**	2.97	7.27	**73.79**	**0.16**	**3.90**	**9.27**	74.54	**0.07**
WANDA	80%	4.55	6.91	65.85	7.57	**4.98**	**9.89**	**66.64**	5.10	4.55	9.87	73.44	1.23
WANDA (matrix)	80%	4.13	6.53	70.10	7.42	17.69	24.21	57.74	8.52	4.31	9.64	73.22	1.64
QK-WANDA	80%	**3.63**	**5.97**	**71.76**	**1.06**	18.98	32.37	66.45	**3.03**	**4.24**	**9.52**	**73.97**	**0.58**

At 80% Q/K sparsity, QK-WANDA substantially narrows the quality gap to the dense Llama 2 70B model (Table 1). WikiText-2/C4 PPL improves from 4.55/6.91 with Wanda to 3.63/5.97, compared with 3.32/5.71 for the dense checkpoint. Mean zero-shot accuracy rises from 65.85% to 71.76%, close to the dense reference of 71.50%. Qwen2.5-72B also improves at this sparsity, although its PPL reductions are smaller (6.7%/3.5%). The contrast is Llama-3.1-70B: QK error decreases from 5.10% to 3.03%, but PPL increases from 4.98/9.89 to 18.98/32.37. This contrast motivates the cross-checkpoint analysis in Section 4.3.

Scoring versus allocation. Matrix-budget Wanda matches separate QK-Wanda's calibration and exact per-projection deletion counts. The following Llama 2 70B comparison at 80% Q/K sparsity separates the two changes:

Scoring and allocation	WikiText-2 PPL	C4 PPL
Wanda, row budgets	4.55	6.91
Wanda, matrix budgets	4.13	6.53
QK-Wanda, separate matrix budgets	3.84	6.14
QK-Wanda, shared budget	3.63	5.97

Both scoring and allocation contributed here. Across the three largest checkpoints and nine sparsities, QK-Wanda with matched matrix budgets had lower QK error in all 27 configurations, lower WikiText-2 PPL in 21, and lower C4 PPL in 20 (Appendix M). These controls establish a scoring benefit on the tested checkpoints, not for every result in the 15-model study. Relaxing Wanda's row quotas also helped in some cases; at 80% on Llama-3.1-70B, matrix-Wanda had lower PPL than either QK allocation, while row-Wanda remained best.

The complementary Llama 2 7B control shows little PPL change when QK-Wanda retains fixed row quotas, but a much larger improvement when deletions can move across rows (Appendix H.2).

Figure 1 gives the full Llama 2 70B sweep. At 50%, the absolute PPL gains were small (WikiText-2/C4 3.36/5.74 to 3.33/5.72). At 90%, PPL decreased from 11.71/13.92 to 4.75/7.01, and QK error from 17.47% to 3.01%; PPL remained above the dense reference of 3.32/5.71. Full family tables are in Appendix F, PPL curves in Appendix H, and QK-error curves in Appendix I.

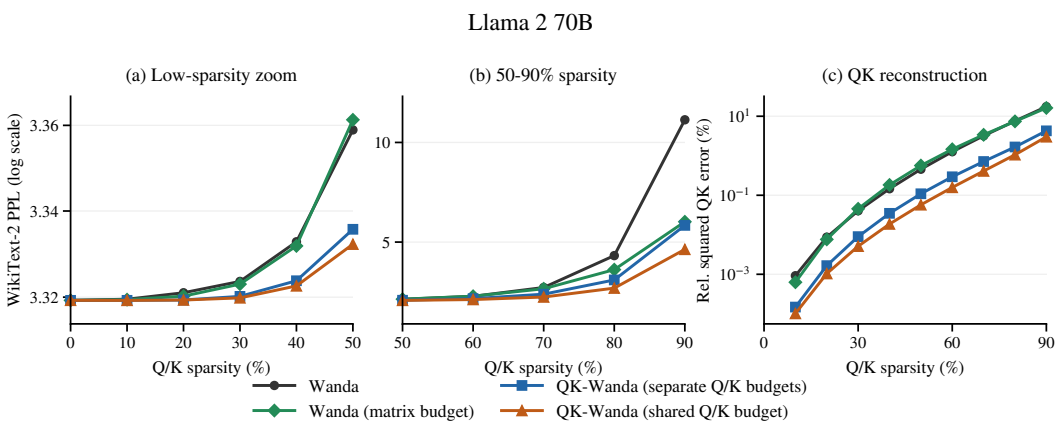

Figure 1: Q/K-only pruning of Llama 2 70B. Matrix-budget Wanda and separate-budget QK-Wanda use identical per-projection deletion counts. Left: low-sparsity zoom of WikiText-2 PPL; center: high-sparsity PPL, with an independent logarithmic range. Right: block-mean relative squared causal pre-RoPE QK error on common dense inputs (Equation (47)); dense zero is omitted. PPL gains are small at moderate sparsity and larger at high sparsity. All non-Q/K weights remain dense. C4: Figure 33; other curves: Appendices H, I.

4.3 Results across all model sizes

Shared-budget QK-Wanda usually improves local reconstruction, but downstream effects depend strongly on checkpoint family and sparsity. Across 135 checkpoint–sparsity configurations, it produced lower measured WikiText-2 PPL in 102, lower C4 PPL in 98, and lower QK error in 133. Mean zero-shot accuracy increased in 27 of 45 evaluated configurations. These sign counts include arbitrarily small differences; Appendix F.4 gives the detailed accounting.

Consistent reconstruction improvements. The mean of the per-checkpoint relative QK-error reductions was 60.8% at 50% sparsity and 45.2% at 80% (Appendix D.3). Nevertheless, lower error can accompany much higher PPL, as Llama-3.1-70B illustrates. The diagnostic evaluates pre-RoPE products on common dense block inputs; it does not measure how RoPE and softmax transform perturbations or how pruning errors accumulate across blocks.

Family-dependent PPL effects. Improvements were most consistent on TinyLlama and Llama 2. At 80%, WikiText-2 PPL fell by 19.2%, 33.6%, and 22.2% on TinyLlama and Llama 2 7B/13B. Qwen2.5 showed small gains at moderate sparsity: across 0.5B–32B, the 50% reductions were 0.22–1.34% on WikiText-2 and 0.09–0.80% on C4. At 80%, Qwen 32B worsened on both datasets, 72B improved on both, and 3B/7B/14B had mixed outcomes. All four Llama 3 sizes had worse PPL on both datasets at 80%. Model size alone therefore did not explain the downstream benefit (Appendix F).

PPL and task accuracy can disagree. At 80%, mean-accuracy gains were 0.84, 5.24, 7.73, and 5.91 percentage points on TinyLlama and Llama 2 7B/13B/70B. Qwen2.5-72B gained 0.52 points, but its 30%/50% means decreased slightly despite lower PPL. Llama-3.1-8B improved mean accuracy at all three evaluated sparsities: at 80%, WikiText-2/C4 PPL worsened from 8.89/13.54 to 13.81/22.36, yet mean accuracy rose from 59.64% to 60.40%. Neither reconstruction error nor PPL alone predicts all task outcomes; Appendix G reports the task-level scores.

4.4 Separate and shared Q/K budgets

Sharing generally lowered measured PPL and QK error relative to separate QK budgets, although the effect varied with model and sparsity (Appendix F.4). We selected shared budgets on the development checkpoint, Llama-3.1-8B, where both PPL metrics improved at 30%/50%/80%, and retained the allocation across checkpoints.

Let n_Q, n_K be the projection weight counts and ρ_Q, ρ_K their removal fractions. A shared budget satisfies

$$\frac{n_Q \rho_Q + n_K \rho_K}{n_Q + n_K} = \rho, \tag{8}$$

up to integer rounding.

Sharing gave larger gains on TinyLlama and Llama 2 70B than on 7B/13B; on 7B at 90%, separate budgets were better (WikiText-2/C4 29.23/40.29 versus 30.14/41.90).

At 50% combined sparsity, mean Q/K removal rates are nearly equal on Llama 2 7B (47.98%/52.02%), but differ substantially on Llama 2 70B (52.74%/28.04%). Both MHA models retain a larger fraction of Q; all 13 GQA models retain a larger fraction of K. These are preferences of the criterion, not an isolated effect of attention architecture. Figure 2 compares the block-level profiles of the three largest checkpoints.

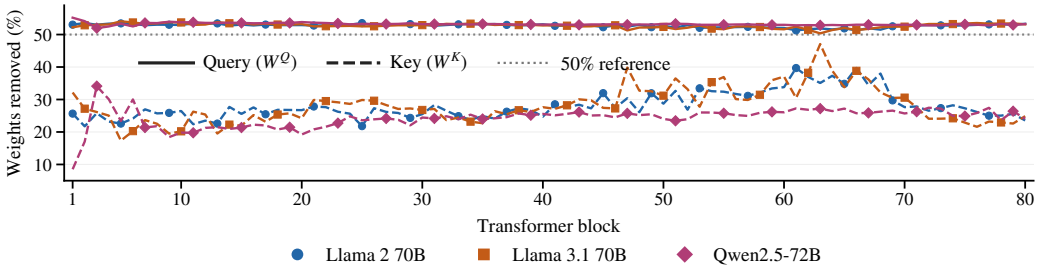

Figure 2: Shared 50% Q/K allocation. Removal rates are per projection; the combined budget is size-weighted. Other model sizes: Figures 3 to 5 (Appendix D.6).

4.5 Pruning all block linear projections

Table 2: Pruning all seven transformer-block projections (Q/K/V/O and MLP gate/up/down), with embeddings and the output head excluded from the sparsity denominator. Wanda uses row budgets throughout; Wanda + QK-Wanda uses shared QK-Wanda only for Q/K and row-wise Wanda elsewhere. Accuracy is the seven-task mean (%); seed 0. Bold marks lower displayed PPL or higher accuracy at each sparsity, including ties.

		TinyLlama-1.1B			**Llama-2-7B**			**Llama-2-13B**		
Method	Sparsity	Wiki2 PPL	C4 PPL	Zero-shot acc. (%)	Wiki2 PPL	C4 PPL	Zero-shot acc. (%)	Wiki2 PPL	C4 PPL	Zero-shot acc. (%)
Dense	0%	7.77	9.94	49.93	5.47	7.26	64.38	4.88	6.73	67.05
Wanda	30%	8.33	10.46	**50.22**	**5.66**	7.50	63.88	5.06	6.90	66.14
Wanda + QK-Wanda	30%	**8.31**	**10.44**	49.94	**5.66**	**7.49**	**64.03**	**5.05**	**6.89**	**66.39**
Wanda	50%	11.52	13.85	48.22	6.89	**9.06**	60.23	5.90	8.00	64.61
Wanda + QK-Wanda	50%	**11.22**	**13.62**	**48.42**	**6.85**	**9.06**	**61.09**	**5.86**	**7.98**	**64.76**
Wanda	80%	512.49	413.72	**34.75**	**1782.14**	**1214.75**	33.76	**1643.05**	**1387.63**	34.65
Wanda + QK-Wanda	80%	**410.73**	**374.82**	34.71	1946.81	2011.70	**34.63**	2037.42	2360.28	**35.24**

To test whether QK-Wanda remains useful as a component of broader pruning, we use it for Q/K while retaining Wanda on V/O and the MLP projections. Table 2 compares this hybrid with Wanda throughout, using the same calibration and evaluation protocol. Here the requested sparsity applies to all seven block projections; embeddings, biases, normalization parameters, and the language-model head remain dense.

At 30% and 50%, the hybrid improved both PPL metrics on all three models (Table 2). At 50%, TinyLlama WikiText-2/C4 PPL decreased from 11.52/13.85 to 11.22/13.62, with smaller relative gains on 7B/13B; mean accuracy increased by 0.19/0.86/0.15 points on TinyLlama/7B/13B. Thus, the Q/K-aware hybrid also helped at moderate full-block sparsity.

At 80% full-block sparsity, both pipelines degraded severely: PPL exceeded 1000 on 7B/13B, and mean accuracy remained below 36% on all three models. Relative differences in this failure regime do not indicate preservation of model quality.

4.6 Calibration, variants, and computational cost

Choice of reconstruction target. Including RoPE in the scoring objective did not consistently improve perplexity. In the 256×32-token, three-seed ablation on Llama 3, the pre-RoPE variant achieved lower mean PPL on both datasets in all six model–sparsity pairs, even though the RoPE-aware variant reduced post-RoPE reconstruction error in four of them (Appendix C). Under the main calibration protocol on Llama 2 7B, pre-RoPE scoring was also better at 40–80% sparsity, whereas RoPE-aware scoring improved both PPL metrics at 90% (Appendix H.2). These results motivate the pre-RoPE default while showing that a more detailed reconstruction target need not yield a better pruning mask.

Progressive re-scoring likewise provided no material PPL benefit in the tested schedules: changes were negligible under row budgets and detrimental under separate matrix budgets (Appendix J). This study and the three-seed ablation used distinct calibration protocols and are excluded from the main aggregate.

Calibration sensitivity. On the development checkpoint, increasing the sample count beyond 256 length-2048 windows produced little additional benefit. At a fixed token budget, length-64 segments gave the lowest QK-Wanda PPL on both datasets, with length 128 close behind. We used the length-64 setting for both methods and retained it across checkpoints (Appendix E).

Computational cost. On Llama 3 1B/3B/8B/70B, the controlled A100 benchmark with $256 \times 32 = 8192$ tokens gave a QK-Wanda/Wanda runtime ratio of 1.31–1.37 for calibration plus scoring and 4.12–4.82 for scoring alone. These measurements excluded loading, transfers, mask selection, and evaluation; these are not end-to-end pruning times at our default calibration budget (Appendix L).

5 Limitations and Scope

Lower QK error guarantees neither lower PPL nor higher accuracy: at 80%, all four Llama 3 checkpoints had worse PPL than Wanda, although three had lower QK error. QK error alone is not a reliable selection rule. Our defaults were developed using the reported WikiText-2/C4 measurements on Llama-3.1-8B, rather than a separate development split. Other checkpoints test transfer of that fixed configuration on the same benchmarks, not untouched benchmark generalization. Future selection studies should reserve separate development and final evaluation data.

The scores are exact for one deletion in the stated causal pre-RoPE objective, not for an optimal many-weight mask or the actual post-RoPE logits. Models with additional nonlinear Q/K transformations require separate treatment. Our checkpoints differ in training, tokenization, and attention architecture; cross-model results do not isolate size or GQA. Unstructured sparsity alone also does not demonstrate inference speedups.

We used one calibration seed for the main comparisons; small differences do not establish statistical significance. Appendix C quantifies three-seed variation under an earlier protocol, which does not bound uncertainty for the main study. We did not study compensating updates to surviving weights under the QK objective.

6 Conclusion

QK-Wanda derives exact single-weight deletion costs for causal pre-RoPE QK reconstruction and constructs masks without gradients or updates to retained weights. Across 15 checkpoints, it generally improves local reconstruction, while downstream benefits depend on checkpoint, sparsity, and metric. Matched-allocation controls show that both scoring and budget allocation can contribute to gains; substantial high-sparsity PPL degradation on Llama 3 demonstrates that improved local reconstruction is insufficient to preserve model quality. Accounting for downstream error propagation and compensating updates are directions for future work.

Ethics Statement

This work studies unstructured pruning through mathematical analysis and experiments on pretrained language models. We used publicly available model checkpoints and established benchmarks, including WikiText-2, C4, and the zero-shot datasets listed in Appendix D.2. We did not recruit human participants or collect new personal or private datasets. Existing models and web-derived corpora can contain biases and sensitive information; pruning does not eliminate these concerns. Compression can facilitate both beneficial uses and misuse of language models, and may change their safety or fairness properties. Our reconstruction and accuracy evaluations do not establish deployment safety; these properties require assessment for the intended application.

Reproducibility Statement

The basic unmasked score derivation appears in Section 3.1; complete formulations, assumptions, and proofs for the causal, unmasked, and RoPE-aware scores are given in Appendices A and B. The pruning procedure is described in Section 3.3. Appendix D specifies data preparation, calibration and evaluation seeds, checkpoint and tokenizer revisions, integer budgets, metrics, software versions, and compute resources. We retained calibration token manifests, pruning masks, source-code manifests, and machine-readable results to support consistency checks and evaluation of the same masks across metrics. An anonymized code-and-results archive is included in the supplementary material; its contents and reproduction instructions are described in Appendix D.5. Executed code is identified by the frozen source manifests there. Calibration sensitivity studies and the controlled timing protocol are reported in Appendices E and L, respectively.

AI Use Statement

We used generative AI tools (ChatGPT/Codex and Claude) to assist with method implementation and debugging, data-preprocessing code, experimental design, refinement of hypotheses, interpretation of results, literature searches, manuscript drafting and editing, and preparation of tables and figures. AI assistance also included formulating and checking mathematical expressions and claims, and developing and checking derivations and proofs. We did not use generative AI to create synthetic benchmark data. All reported numerical results were obtained by executing the pruning and evaluation procedures described in this paper. The human authors reviewed and verified the AI-assisted code, text, mathematical reasoning, and empirical claims, and take full responsibility for the final content of this work.

REFERENCES

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*, pp. 4895–4901, 2023. URL https://aclanthology.org/2023.emnlp-main.298/.

Saleh Ashkboos, Maximilian L. Croci, Marcelo Gennari do Nascimento, Torsten Hoefler, and James Hensman. SliceGPT: Compress large language models by deleting rows and columns. In *International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=vXxardq6db.

Christopher Clark, Kenton Lee, Ming-Wei Chang, Tom Kwiatkowski, Michael Collins, and Kristina Toutanova. BoolQ: Exploring the surprising difficulty of natural yes/no questions. In *Proceedings of NAACL-HLT*, 2019.

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

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

EleutherAI. Eleutherai lm-evaluation-harness v0.4.12. Zenodo, 2026.

Elias Frantar and Dan Alistarh. SparseGPT: Massive language models can be accurately pruned in one-shot. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 10323–10337. PMLR, 2023. URL https://proceedings.mlr.press/v202/frantar23a.html.

Elias Frantar, Sidak Pal Singh, and Dan Alistarh. Optimal brain compression: A framework for accurate post-training quantization and pruning. In *Advances in Neural Information Processing Systems*, volume 35, 2022. URL https://arxiv.org/abs/2208.11580.

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.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024. URL https://arxiv.org/abs/2407.21783.

Babak Hassibi and David G. Stork. Second order derivatives for network pruning: Optimal Brain Surgeon. In *Advances in Neural Information Processing Systems*, volume 5, 1992. URL https://proceedings.neurips.cc/paper/1992/hash/303ed4c69846ab36c2904d3ba8573050-Abstract.html.

Yang He and Joey Tianyi Zhou. Data-independent module-aware pruning for hierarchical vision transformers. In *International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=7Ol6foUi1G.

Torsten Hoefler, Dan Alistarh, Tal Ben-Nun, Nikoli Dryden, and Alexandra Peste. Sparsity in deep learning: Pruning and growth for efficient inference and training in neural networks. *Journal of Machine Learning Research*, 22(241):1–124, 2021. URL https://jmlr.org/papers/v22/21-0366.html.

Hanyu Hu, Pengxiang Zhao, Ping Li, Yi Zheng, Zhefeng Wang, and Xiaoming Yuan. FASP: Fast and accurate structured pruning of large language models. *arXiv preprint arXiv:2501.09412*, 2025. URL https://arxiv.org/abs/2501.09412.

Ivan Ilin and Peter Richtárik. Thanos: A block-wise pruning algorithm for efficient large language model compression. *arXiv preprint arXiv:2504.05346*, 2025. URL https://arxiv.org/abs/2504.05346.

Yixin Ji, Yang Xiang, Juntao Li, Qingrong Xia, Ping Li, Xinyu Duan, Zhefeng Wang, and Min Zhang. Beware of calibration data for pruning large language models. In *International Conference on Learning Representations*, 2025. URL https://arxiv.org/abs/2410.17711.

Samir Khaki, Xiuyu Li, Junxian Guo, Ligeng Zhu, Konstantinos N. Plataniotis, Amir Yazdanbakhsh, Kurt Keutzer, Song Han, and Zhijian Liu. SparseLoRA: Accelerating LLM fine-tuning with contextual sparsity. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 29768–29783. PMLR, 2025. URL https://proceedings.mlr.press/v267/khaki25a.html.

Junhan Kim, Chungman Lee, Eulrang Cho, Kyungphil Park, Ho-young Kim, Joonyoung Kim, and Yongkweon Jeon. Towards next-level post-training quantization of hyper-scale transformers. In *Advances in Neural Information Processing Systems*, volume 37, 2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/ab6eba9a853087993addff937c8cec87-Abstract-Conference.html.

Junhan Kim, Ho-Young Kim, Eulrang Cho, Chungman Lee, Joonyoung Kim, and Yongkweon Jeon. BoA: Attention-aware post-training quantization without backpropagation. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 30132–30152. PMLR, 2025. URL https://proceedings.mlr.press/v267/kim25j.html.

Yann LeCun, John S. Denker, and Sara A. Solla. Optimal brain damage. In *Advances in Neural Information Processing Systems*, volume 2, 1989. URL https://proceedings.neurips.cc/paper/1989/hash/6c9882bbac1c7093bd25041881277658-Abstract.html.

Yingyu Liang, Jiangxuan Long, Zhenmei Shi, Zhao Song, and Yufa Zhou. Beyond linear approximations: A novel pruning approach for attention matrix. In *International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=sgbI8Pxwie.

Xinyin Ma, Gongfan Fang, and Xinchao Wang. LLM-Pruner: On the structural pruning of large language models. In *Advances in Neural Information Processing Systems*, volume 36, 2023. URL https://arxiv.org/abs/2305.11627.

Vladimir Malinovskii, Denis Mazur, Ivan Ilin, Denis Kuznedelev, Konstantin Burlachenko, Kai Yi, Dan Alistarh, and Peter Richtárik. PV-Tuning: Beyond straight-through estimation for extreme LLM compression. In *Advances in Neural Information Processing Systems*, volume 37, 2024a. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/091166620a04a289c555f411d8899049-Abstract-Conference.html.

Vladimir Malinovskii, Andrei Panferov, Ivan Ilin, Han Guo, Peter Richtárik, and Dan Alistarh. Pushing the limits of large language model quantization via the linearity theorem. *arXiv preprint arXiv:2411.17525*, 2024b. URL https://arxiv.org/abs/2411.17525.

Fanxu Meng, Pingzhi Tang, Fan Jiang, and Muhan Zhang. CLOVER: Cross-layer orthogonal vectors pruning. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 43807–43825. PMLR, 2025. URL https://proceedings.mlr.press/v267/meng25d.html.

Stephen Merity, Caiming Xiong, James Bradbury, and Richard Socher. Pointer sentinel mixture models. In *International Conference on Learning Representations*, 2017. URL https://arxiv.org/abs/1609.07843.

Meta AI. Llama 3.2 model card. Official model documentation, 2024. URL https://huggingface.co/meta-llama/Llama-3.2-1B.

Todor Mihaylov, Peter Clark, Tushar Khot, and Ashish Sabharwal. Can a suit of armor conduct electricity? a new dataset for open book question answering. In *Proceedings of EMNLP*, 2018.

Qwen, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, et al. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*, 2024. URL https://arxiv.org/abs/2412.15115.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67, 2020. URL https://jmlr.org/papers/v21/20-074.html.

Keisuke Sakaguchi, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. WinoGrande: An adversarial winograd schema challenge at scale. In *Proceedings of AAAI*, 2020.

Kyunghwan Shim, Jaewoong Yun, and Shinkook Choi. SNP: Structured neuron-level pruning to preserve attention scores. In *European Conference on Computer Vision*, 2024. URL https://www.ecva.net/papers/eccv_2024/papers_ECCV/html/10737_ECCV_2024_paper.php.

Jianlin Su, Murtadha Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. RoFormer: Enhanced transformer with rotary position embedding. *Neurocomputing*, 568:127063, 2024. doi: 10.1016/j.neucom.2023.127063.

Mingjie Sun, Zhuang Liu, Anna Bair, and J. Zico Kolter. A simple and effective pruning approach for large language models. In *International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=PxoFut3dWW.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023. URL https://arxiv.org/abs/2307.09288.

Albert Tseng, Qingyao Sun, David Hou, and Christopher De Sa. QTIP: Quantization with trellises and incoherence processing. In *Advances in Neural Information Processing Systems*, volume 37, 2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/6de2e84b8da47bb2eb5e2ac96c63d2b0-Abstract-Conference.html.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30, 2017. URL https://arxiv.org/abs/1706.03762.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *Proceedings of the EMNLP Workshop BlackboxNLP*, 2018.

Jeffrey T. H. Wong, Cheng Zhang, Xinye Cao, Pedro Gimenes, Christos-Savvas Bouganis, George A. Constantinides, Wayne Luk, and Yiren Zhao. A^3: An analytical low-rank approximation framework for attention. *arXiv preprint arXiv:2505.12942*, 2025. URL https://arxiv.org/abs/2505.12942v4. Version 4, revised May 2026.

Jihao Xin, Tian Lyu, David Keyes, Hatem Ltaief, and Marco Canini. RAP: KV-cache compression via RoPE-aligned pruning. *arXiv preprint arXiv:2602.02599*, 2026. URL https://arxiv.org/abs/2602.02599.

Kai Yi and Peter Richtárik. Symmetric pruning of large language models. *arXiv preprint arXiv:2501.18980*, 2025. URL https://arxiv.org/abs/2501.18980.

Lu Yin, You Wu, Zhenyu Zhang, Cheng-Yu Hsieh, Yaqing Wang, Yiling Jia, Gen Li, Ajay Kumar Jaiswal, Mykola Pechenizkiy, Yi Liang, Michael Bendersky, Zhangyang Wang, and Shiwei Liu. Outlier weighed layerwise sparsity (OWL): A missing secret sauce for pruning LLMs to high sparsity. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 57101–57115. PMLR, 2024. URL https://proceedings.mlr.press/v235/yin24e.html.

Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. HellaSwag: Can a machine really finish your sentence? In *Proceedings of ACL*, 2019.

Boxiang Zhang and Baijian Yang. CORP: Closed-form one-shot representation-preserving structured pruning for transformers. *arXiv preprint arXiv:2602.05243*, 2026. URL https://arxiv.org/abs/2602.05243v2. Version 2, revised May 2026.

Peiyuan Zhang, Guangtao Zeng, Tianduo Wang, and Wei Lu. TinyLlama: An open-source small language model. *arXiv preprint arXiv:2401.02385*, 2024. URL https://arxiv.org/abs/2401.02385.

A Complete Proofs for the Causal Score

Appendix roadmap. Appendices A–B contain the score derivations; D documents data and implementation, and E covers default selection. Full family tables, task scores, and curves are in Appendices F–I. Controlled variants and row budgets are in Appendices C and H.2; progressive re-scoring, timing, and matched matrix budgets are in Appendices J, L, and M.

A.1 Full objective, general scores, and their proof

We extend Section 3.2 to multiple calibration sequences and shared-key heads. All identities hold for fixed inputs and real-valued arithmetic in one transformer block. The calibration set contains N token sequences, indexed by $s = 1, \dots, N$. Sequence s has T_s tokens, whose input activations form $X^{(s)} \in \mathbb{R}^{d \times T_s}$: a real-valued matrix with d features and T_s columns. Parenthesized superscripts such as (s) identify a sequence, not an exponent. We take the arithmetic mean of the sequence losses, so each sequence contributes with factor $1/N$. This matches our experiments and requires no sequence-importance parameters.

Heads and coordinates. Let H_Q and H_K be the numbers of query and key heads, and let d_h be their common number of coordinates per head, corresponding to m in the single-head presentation of Section 3.1. A query head has index $h \in \{1, \dots, H_Q\}$; a key head has index $g \in \{1, \dots, H_K\}$. The function $g(h)$ identifies the key head used by query head h. The set $\mathcal{H}_g = \{h : g(h) = g\}$ contains all query heads that use key head g. This covers grouped-query attention (GQA): for example, four query heads can share two key heads, with $\mathcal{H}_1 = \{1, 2\}$ and $\mathcal{H}_2 = \{3, 4\}$. In ordinary multi-head attention, $g(h) = h$ and each set contains one head. All heads within a projection share the same width, so omitting the attention scale $1/\sqrt{d_h}$ multiplies every squared deletion score by the same positive constant d_h. The Q/K score ordering is unchanged for both separate and shared budgets.

The matrices $W^{Q,h}, W^{K,g} \in \mathbb{R}^{d_h \times d}$ map input features to head coordinates; their superscripts specify projection type and head. We write their scalar entries as $W^{Q,h}_{ij}$ and $W^{K,g}_{ij}$, where $i = 1, \dots, d_h$ and $j = 1, \dots, d$. The same coordinate i is paired in each dot product: all heads have width d_h, so changing $K_{g,iu}$ affects $Q_{h,it}$ for every $h \in \mathcal{H}_g$. The optional bias vectors $b^{Q,h}, b^{K,g}$ have d_h entries. Let $e^{(s)} \in \{1\}^{T_s \times 1}$ be the all-ones column vector for sequence s, using the transpose notation from Section 3.1. Before rotary position embedding (RoPE), the activations are

$$Q^{(s)}_h = W^{Q,h} X^{(s)} + b^{Q,h} (e^{(s)})^\top, \qquad K^{(s)}_g = W^{K,g} X^{(s)} + b^{K,g} (e^{(s)})^\top. \tag{9}$$

Both activation matrices have d_h rows and T_s columns. In entry notation, $Q^{(s)}_{h,it}$ is query coordinate i at token t, and $K^{(s)}_{g,iu}$ is key coordinate i at token u. Biases are retained during pruning; for models without biases, the bias vectors are zero. A fixed bias cancels in the change of its own projection, but remains in the opposite projection used by the score. It therefore cannot generally be discarded as an additive constant in the reconstruction loss.

Which token pairs contribute? Use t for a query position and u for a key position, both between 1 and T_s. Define the $T_s \times T_s$ matrices

$$B^{(s)}_h = Q^{(s)\top}_h K^{(s)}_{g(h)}, \qquad C^{(s)}_{tu} = \begin{cases} 1, & u \le t, \\ 0, & u > t. \end{cases} \tag{10}$$

The entry $B^{(s)}_{h,tu}$ sums the products of matching coordinates of query token t and key token u, following the unscaled convention in Section 3.1. The binary matrix $C^{(s)}$ is the causal mask: a query can use its current and preceding tokens, so only $u \le t$ contributes. For three tokens it is $C = \left(\begin{smallmatrix} 1 & 0 & 0 \\ 1 & 1 & 0 \\ 1 & 1 & 1 \end{smallmatrix}\right)$.

The reconstruction loss. The hatted product $\widehat{B}^{(s)}_h$ denotes the value after pruning and is compared with its fixed reference $B^{(s)}_h$. For an error matrix E, $C \odot E$ denotes entrywise multiplication, $(C \odot E)_{tu} = C_{tu} E_{tu}$. It keeps allowed entries and sets forbidden entries to zero. We use the squared

Frobenius norm defined in Section 3.1. The subscript C distinguishes this causal loss from the unmasked $\mathcal{L}$ in Equation (3). With these definitions,

$$\mathcal{L}_C = \frac{1}{N}\sum_{s=1}^{N}\sum_{h=1}^{H_Q}\left\|C^{(s)} \odot \left(\widehat{B}_h^{(s)} - B_h^{(s)}\right)\right\|_F^2$$
$$= \frac{1}{N}\sum_{s=1}^{N}\sum_{h=1}^{H_Q}\sum_{t=1}^{T_s}\sum_{u=1}^{t}\left(\widehat{B}_{h,tu}^{(s)} - B_{h,tu}^{(s)}\right)^2. \tag{11}$$

In words: square the change of each allowed query–key dot product, add over token pairs and heads, then average over calibration sequences. Independent sequences are evaluated separately; no dot product connects tokens from different sequences. The mask selects entries rather than adding the negative infinities used inside softmax. Using a sum instead of the average would multiply all scores by the same positive number N, leaving their ranking unchanged.

Let $\hat{S}_{ij}^{Q,h}$ denote the value of $\mathcal{L}_C$ after setting the query weight $W_{ij}^{Q,h}$ to zero while every other weight stays fixed. Define $\hat{S}_{ij}^{K,g}$ in the same way for the key weight $W_{ij}^{K,g}$. As in the main text, a hat on S labels the causal score. These scores measure the damage from each individual deletion; smaller is better to prune.

Proposition A.1 (Exact causal single-weight deletion)**.** *Fix the calibration inputs, projections, and retained biases defined above. The query score is*

$$\boxed{\hat{S}_{ij}^{Q,h} = \frac{(W_{ij}^{Q,h})^2}{N}\sum_{s=1}^{N}\sum_{t=1}^{T_s}(X_{jt}^{(s)})^2\sum_{u=1}^{t}(K_{g(h),iu}^{(s)})^2.} \tag{12}$$

For a key weight shared by the query heads in $\mathcal{H}_g$, it is

$$\boxed{\hat{S}_{ij}^{K,g} = \frac{(W_{ij}^{K,g})^2}{N}\sum_{s=1}^{N}\sum_{u=1}^{T_s}(X_{ju}^{(s)})^2\sum_{h\in\mathcal{H}_g}\sum_{t=u}^{T_s}(Q_{h,it}^{(s)})^2.} \tag{13}$$

Setting $N = 1$ and using one query head per key head reduces these expressions to Equation (7): the inner sums are precisely the squared prefix/suffix norms. The factor $1/N$ averages independent sequence losses; it does not introduce token pairs between different sequences.

The reference pair has zero reconstruction loss. These scores measure the loss after one finite deletion, not language-model cross-entropy. The proof first allows any fixed binary mask $C^{(s)}$, then specializes to the causal mask. Sums over t, u range from 1 to T_s unless a causal restriction is stated.

Proof of Proposition A.1. For a weight, activation, or product matrix M, write $\Delta M = \widehat{M} - M$ for its pruned value minus its reference value. Let $e_i \in \mathbb{R}^{d_h}$ be the vector that is one at coordinate i and zero elsewhere, and define $e_j \in \mathbb{R}^d$ similarly. These are standard basis vectors, distinct from the all-ones vector $e^{(s)}$. The outer product $e_i e_j^\top$ selects exactly one weight coordinate.

Query deletion. For $w = W_{ij}^{Q,h}$, setting this weight to zero means

$$\Delta W^{Q,h} = -we_ie_j^\top, \qquad \Delta Q_h^{(s)} = -we_iX_{j:}^{(s)}. \tag{14}$$

All keys and other query heads remain fixed. The bias cancels when taking the difference of the two affine projections. Hence

$$\widehat{B}_h^{(s)} - B_h^{(s)} = (Q_h^{(s)} + \Delta Q_h^{(s)})^\top K_{g(h)}^{(s)} - Q_h^{(s)\top}K_{g(h)}^{(s)}$$
$$= -wX_{j:}^{(s)\top}K_{g(h),i:}^{(s)}. \tag{15}$$

In particular, the (t, u) entry is exactly

$$\Delta B_{h,tu}^{(s)} = -wX_{jt}^{(s)}K_{g(h),iu}^{(s)}. \tag{16}$$

Expanding the Frobenius norm entry by entry gives

$$\mathcal{L}_C = \frac{1}{N}\sum_{s=1}^{N}\sum_{t,u}\big(C_{tu}^{(s)}\Delta B_{h,tu}^{(s)}\big)^2$$
$$= \frac{w^2}{N}\sum_{s=1}^{N}\sum_{t,u} C_{tu}^{(s)}(X_{jt}^{(s)})^2(K_{g(h),iu}^{(s)})^2. \tag{17}$$

We used $C_{tu}^2 = C_{tu}$, which holds because every mask entry is zero or one; all other heads contribute zero. For the causal mask, only $u \le t$ remains in the inner sum, yielding Equation (12).

Key deletion and GQA. For $w = W_{ij}^{K,g}$, the corresponding finite change is

$$\Delta W^{K,g} = -we_i e_j^\top, \qquad \Delta K_g^{(s)} = -we_i X_{j:}^{(s)}. \tag{18}$$

For each $h \in \mathcal{H}_g$, queries remain fixed, so

$$\widehat{B}_h^{(s)} - B_h^{(s)} = Q_h^{(s)\top}\Delta K_g^{(s)} = -wQ_{h,i:}^{(s)\top}X_{j:}^{(s)}. \tag{19}$$

The entry change is $-wQ_{h,it}^{(s)}X_{ju}^{(s)}$. Heads outside $\mathcal{H}_g$ do not change. Therefore

$$\mathcal{L}_C = \frac{1}{N}\sum_{s=1}^{N}\sum_{h\in\mathcal{H}_g}\sum_{t,u}\big(C_{tu}^{(s)}\Delta B_{h,tu}^{(s)}\big)^2$$
$$= \frac{w^2}{N}\sum_{s=1}^{N}\sum_{u}(X_{ju}^{(s)})^2\sum_{h\in\mathcal{H}_g}\sum_{t} C_{tu}^{(s)}(Q_{h,it}^{(s)})^2. \tag{20}$$

The loss sums errors in distinct head matrices, so there are no cross-products between query heads. For the causal mask, valid query positions satisfy $t \ge u$, giving Equation (13). Both derivations are finite algebraic identities, with no Taylor truncation. Biases contribute to the fixed opposite-side activations throughout. □

A general binary mask can also express padding, sliding windows, or document boundaries. The evaluated runner uses fixed-length unpadded windows with ordinary causal attention; the mathematical statement does not imply implementation support for every mask layout.

A.2 A TWO-TOKEN EXAMPLE

Take one sequence and one head, with $N = 1$, $T = 2$, $d = 2$, $d_h = 1$, and zero biases. Suppress the sequence/head labels. Let

$$X = \begin{pmatrix}1 & 2\\ 3 & 4\end{pmatrix}, \qquad W^Q = \begin{pmatrix}2 & 1\end{pmatrix}, \qquad W^K = \begin{pmatrix}1 & 0\end{pmatrix}.$$

Then $Q = (5, 8)$ and $K = (1, 2)$. Delete $W_{11}^Q = 2$, the weight from input feature 1 to query coordinate 1. The pruned query is $\widehat{Q} = (3, 4)$. The product change and its causal restriction are

$$\widehat{B} - B = \begin{pmatrix}-2 & -4\\ -4 & -8\end{pmatrix}, \qquad C \odot (\widehat{B} - B) = \begin{pmatrix}-2 & 0\\ -4 & -8\end{pmatrix}.$$

The top-right entry is ignored because query token 1 cannot attend to key token 2. The loss is $\mathcal{L}_C = (-2)^2 + (-4)^2 + (-8)^2 = 84$. The closed-form score gives the same answer:

$$\hat{S}_{11}^Q = 2^2\big[1^2(1^2) + 2^2(1^2 + 2^2)\big] = 84.$$

The two terms in brackets use the key prefixes available to query tokens 1 and 2. This example makes explicit what the mask, squared norm, and prefix sum each contribute.

A.3 When do row-wise masks coincide with Wanda?

For one unmasked sequence, Equation (5) multiplies every Wanda score in query row i by the same factor $\|K_{i:}\|_2^2$, and every score in key row i by $\|Q_{i:}\|_2^2$. A positive factor preserves rankings and ties, so identical per-row deletion counts and tie handling give identical masks. If the factor is zero, all QK scores in that row tie and equivalence is not guaranteed. The same argument holds for the grouped opposite-head energy in the single-sequence GQA formula (Section B.1). For multiple independent sequences, the exact score averages products of input and opposite-side energies; these generally cannot be factored into a Wanda score times a row constant.

Causal masking can break the equivalence even for one sequence and one head. Set

$$X = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}, \qquad W^Q = (2 \quad 1), \qquad W^K = (1 \quad 3),$$

with zero biases. Then $Q = (2, 1)$ and $K = (1, 3)$. For the two query weights, Wanda gives scores $(4, 1)$, and the unmasked score gives $(40, 10)$: both delete the second weight at 50% row sparsity. The causal scores are instead $(4, 10)$, because the first query sees only the first key whereas the second sees both. They delete the first weight. Thus a row-wise causal control is informative; the unmasked equivalence is not a reason to omit it.

A.4 Derivation of the prefix/suffix accumulators

For sequence s, define two $d_h \times T_s$ arrays. The notation $[i, t]$ selects the entry at row i, column t:

$$P_K^{(g,s)}[i,t] = \sum_{u \le t} (K_{g,iu}^{(s)})^2, \tag{21}$$

$$U_Q^{(g,s)}[i,u] = \sum_{h \in \mathcal{H}_g} \sum_{t \ge u} (Q_{h,it}^{(s)})^2. \tag{22}$$

The array P_K stores key-prefix sums; U_Q stores grouped-query-suffix sums. They are computed by forward and reverse cumulative sums. Let $X_2^{(s)} = X^{(s)} \odot X^{(s)}$ denote the input matrix with each entry squared. For any matrix M, $[M]_{ij}$ also denotes its row-i, column-j entry. Thus

$$[P_K^{(g,s)} X_2^{(s)\top}]_{ij} = \sum_t \sum_{u \le t} (K_{g,iu}^{(s)})^2 (X_{jt}^{(s)})^2, \tag{23}$$

$$[U_Q^{(g,s)} X_2^{(s)\top}]_{ij} = \sum_u \sum_{h \in \mathcal{H}_g} \sum_{t \ge u} (Q_{h,it}^{(s)})^2 (X_{ju}^{(s)})^2. \tag{24}$$

Multiplying by the corresponding squared weight, then averaging over the N sequences, gives exactly Equations (12) and (13). Each resulting interaction matrix has shape $d_h \times d$. Query factors are reused within a GQA group; their weight multipliers differ by head. These are exact rearrangements of the pairwise sums.

The implementation accumulates in FP32. Within a block, both masks are formed before either is applied, then the sparse block output is propagated. Thus the inputs are fixed during a local scoring decision, even though later blocks receive method-specific sparse inputs.

A.5 Quadratic form and its diagonal

The notation $X_{:t}$ selects column t, so write the input, key, and query column vectors as $x_t^{(s)} = X_{:t}^{(s)}$, $k_{g,u}^{(s)} = K_{g,:u}^{(s)}$, and $q_{h,t}^{(s)} = Q_{h,:t}^{(s)}$. The operation $\mathrm{vec}(M)$ stacks the columns of a matrix M into one column vector, in order. The Kronecker product $A \otimes V$ is the block matrix whose block in position (i, j) is $A_{ij}V$. This differs from the entrywise product $\odot$.

For a change to one query-head matrix, let $\delta w = \mathrm{vec}(\Delta W^{Q,h}) \in \mathbb{R}^{dd_h}$. With keys fixed, the change of an unscaled dot product can be written as

$$k_{g(h),u}^{(s)\top} \Delta W^{Q,h} x_t^{(s)} = \left(x_t^{(s)} \otimes k_{g(h),u}^{(s)}\right)^\top \mathrm{vec}(\Delta W^{Q,h}). \tag{25}$$

Squaring this scalar and summing valid entries proves that the exact fixed-key reconstruction loss is $\delta w^\top G^{Q,h} \delta w$, where

$$G^{Q,h} = \frac{1}{N} \sum_{s=1}^{N} \sum_{t,u} C_{tu}^{(s)} (x_t^{(s)} x_t^{(s)\top}) \otimes (k_{g(h),u}^{(s)} k_{g(h),u}^{(s)\top}). \tag{26}$$

The matrix $G^{Q,h} \in \mathbb{R}^{dd_h \times dd_h}$ therefore collects the coefficients of this quadratic loss. For a change to key head g, use $\delta w = \mathrm{vec}(\Delta W^{K,g})$ instead. Keeping queries fixed and including all affected heads gives the coefficient

$$G^{K,g} = \frac{1}{N} \sum_{s=1}^{N} \sum_{u} \sum_{h \in \mathcal{H}_g} \sum_{t} C_{tu}^{(s)} (x_u^{(s)} x_u^{(s)\top}) \otimes (q_{h,t}^{(s)} q_{h,t}^{(s)\top}). \tag{27}$$

Write G for either coefficient matrix when the side/head is clear. Each is positive semidefinite: $v^\top G v \geq 0$ for every vector v, because it represents a sum of squared dot-product changes. The Hessian, the matrix of second derivatives with respect to δw, is $2G$. Matrix entry (i, j) has vectorized index $p = i + (j-1)d_h$. Let w_p be that vector entry and $e_p \in \mathbb{R}^{dd_h}$ select it. Deleting it means $\delta w = -w_p e_p$, yielding $\hat{S}_p = w_p^2 [G]_{pp}$. The diagonal coefficient $[G]_{pp}$ determines the cost of perturbing coordinate p; deleting that weight incurs the exact cost $w_p^2 [G]_{pp}$. Off-diagonal entries describe interactions between weights. The causal restriction generally prevents factorization into separately aggregated input and opposite-projection Gram matrices (products such as $XX^\top$).

A.6 Scaling and the limits of scalar exactness

Multiply all query weights and biases in one GQA group by a nonzero scalar c, and the shared key weights and bias by its reciprocal $c^{-1} = 1/c$. In each Q score, the squared query weight gains c^2 and key energy gains c^{-2}; the factors cancel. The key score has the reciprocal cancellation. Thus products and QK-Wanda scores are invariant. Including biases is required for complete activations to scale this way; groupwise scalar rescaling also commutes with RoPE.

For several query deletions, $\delta w^\top G \delta w$ contains off-diagonal interactions. Summing scalar scores retains only its diagonal terms. For one sequence and one head, suppress their labels. If both projections change, exact expansion gives

$$\widehat{Q}^\top \widehat{K} - Q^\top K = \Delta Q^\top K + Q^\top \Delta K + \Delta Q^\top \Delta K. \tag{28}$$

Apply the causal mask to this complete change. Its squared norm also contains cross terms between the two linear paths and the bilinear path. Independent scalar thresholding therefore does not give the exact joint loss or an optimal full mask.

B Proofs for the Unmasked and RoPE-Aware Variants

B.1 QK-Wanda-U: multiple sequences and shared keys

Section 3.1 derives the unmasked scores for one sequence and one head. Here we state the general form using the calibration and head notation from Section A.1.

Corollary B.1 (Exact unmasked scalar scores)**.** *Under the assumptions of Proposition A.1, replace the causal mask with $C_{tu}^{(s)} = 1$ for every token pair. The resulting loss $\mathcal{L}$ is the sequence average of Equation* (3)*. Unmasked scores are written as S without a hat. The single-weight costs are*

$$S_{ij}^{Q,h} = \frac{(W_{ij}^{Q,h})^2}{N} \sum_{s=1}^{N} \left\| X_{j:}^{(s)} \right\|_2^2 \left\| K_{g(h),i:}^{(s)} \right\|_2^2, \tag{29}$$

$$S_{ij}^{K,g} = \frac{(W_{ij}^{K,g})^2}{N} \sum_{s=1}^{N} \left\| X_{j:}^{(s)} \right\|_2^2 \sum_{h \in \mathcal{H}_g} \left\| Q_{h,i:}^{(s)} \right\|_2^2. \tag{30}$$

Proof. Set $C_{tu}^{(s)} = 1$ in Equations (17) and (20). The sum over input positions separates from the sum over opposite-projection positions within each sequence. Equivalently, the query perturbation in

Equation (15) is an outer product. For any vectors u, v,

$$\left\|uv^\top\right\|_F^2 = \sum_{t,r} u_t^2 v_r^2 = \left(\sum_t u_t^2\right)\left(\sum_r v_r^2\right). \tag{31}$$

Applying this identity to the query perturbation gives Equation (29). For a key deletion, apply it to $Q_{h,i:}^\top X_{j:}$ in each affected head and sum the squared losses over $\mathcal{H}_g$, giving Equation (30). □

For one sequence ($N = 1$) and one query head per key head, suppress sequence/head labels. The scores reduce to

$$S_{ij}^Q = (W_{ij}^Q)^2 \left\|X_{j:}\right\|_2^2 \left\|K_{i:}\right\|_2^2, \qquad S_{ij}^K = (W_{ij}^K)^2 \left\|X_{j:}\right\|_2^2 \left\|Q_{i:}\right\|_2^2. \tag{32}$$

Within a fixed output row, the opposite-side factor is independent of j. If positive, it preserves the ordering of WANDA scores; equal row budgets and the same tie convention then give identical masks. The argument also holds for the grouped-query energy in a shared key row. A zero factor gives a degenerate tie. Matrix-global allocation may change the budget across rows; multiple sequences and causal weighting generally do not preserve this row-ranking equivalence.

For QK-WANDA-U, summing all pairs in Equation (26) yields

$$G_U^{Q,h} = \frac{1}{N}\sum_{s=1}^{N} (X^{(s)} X^{(s)\top}) \otimes (K_{g(h)}^{(s)} K_{g(h)}^{(s)\top}). \tag{33}$$

The factorization is sequence-local; a sum of products need not equal a product of sums.

B.2 QK-WANDA-R: CAUSAL RECONSTRUCTION INCLUDING ROTARY EMBEDDINGS

The column vectors $q_{h,t}^{(s)}$ and $k_{g,u}^{(s)}$, defined in Section A.5, contain all d_h coordinates of one query/key token. Let $R_t^{(s)} \in \mathbb{R}^{d_h \times d_h}$ be the model's fixed rotary position map for token t in sequence s. A tilde marks an activation after this map: $\widetilde{q}_{h,t}^{(s)} = R_t^{(s)} q_{h,t}^{(s)}$ and $\widetilde{k}_{g,u}^{(s)} = R_u^{(s)} k_{g,u}^{(s)}$. The same maps are used for the reference and pruned weights. Let $Z_h^{(s)}$ be the $T_s \times T_s$ matrix of unscaled dot products *after* RoPE; this distinguishes it from the pre-RoPE matrix $B_h^{(s)}$. Its entries and reconstruction loss are

$$Z_{h,tu}^{(s)} = \widetilde{q}_{h,t}^{(s)\top} \widetilde{k}_{g(h),u}^{(s)}, \qquad \mathcal{L}_R = \frac{1}{N}\sum_{s=1}^{N}\sum_h \left\| C^{(s)} \odot (\widehat{Z}_h^{(s)} - Z_h^{(s)}) \right\|_F^2. \tag{34}$$

The loss $\mathcal{L}_R$ uses the same unscaled convention as $\mathcal{L}_C$. Here C retains causal validity; QK-WANDA-R adds the rotary transform rather than replacing the causal mask.

Proposition B.2 (Exact RoPE-aware single-weight deletion)**.** *Under the linear Q/K assumptions of Proposition A.1, and fixed rotary maps as above, add the suffix R to the causal scores $\hat{S}$ to indicate that RoPE is included. With e_i the basis vector selecting raw coordinate i, the exact single-deletion costs in $\mathcal{L}_R$ are*

$$\hat{S}_{ij,R}^{Q,h} = \frac{(W_{ij}^{Q,h})^2}{N}\sum_{s=1}^{N}\sum_{t,u} C_{tu}^{(s)} (X_{jt}^{(s)})^2 \left[(R_t^{(s)} e_i)^\top \widetilde{k}_{g(h),u}^{(s)}\right]^2, \tag{35}$$

$$\hat{S}_{ij,R}^{K,g} = \frac{(W_{ij}^{K,g})^2}{N}\sum_{s=1}^{N}\sum_{u} (X_{ju}^{(s)})^2 \sum_{h\in\mathcal{H}_g}\sum_t C_{tu}^{(s)} \left[\widetilde{q}_{h,t}^{(s)\top} R_u^{(s)} e_i\right]^2. \tag{36}$$

Proof. Deleting a raw query weight $w = W_{ij}^{Q,h}$ gives $\Delta q_{h,t}^{(s)} = -w X_{jt}^{(s)} e_i$. Since $R_t^{(s)}$ is fixed and linear,

$$\Delta\widetilde{q}_{h,t}^{(s)} = R_t^{(s)} \Delta q_{h,t}^{(s)} = -w X_{jt}^{(s)} R_t^{(s)} e_i. \tag{37}$$

Keys remain fixed. Subtracting the reference dot product therefore yields

$$\Delta Z_{h,tu}^{(s)} = -w X_{jt}^{(s)} (R_t^{(s)} e_i)^\top \widetilde{k}_{g(h),u}^{(s)}. \tag{38}$$

Expand $\mathcal{L}_R$ as $\frac{1}{N}\sum_{s=1}^{N}\sum_{t,u}(C_{tu}^{(s)}\Delta Z_{h,tu}^{(s)})^2$. Use $C_{tu}^2 = C_{tu}$, and substitute the preceding identity to obtain Equation (35) exactly.

For a raw key deletion $w = W_{ij}^{K,g}$,

$$\Delta\widetilde{k}_{g,u}^{(s)} = -wX_{ju}^{(s)}R_u^{(s)}e_i, \qquad \Delta Z_{h,tu}^{(s)} = -wX_{ju}^{(s)}\widetilde{q}_{h,t}^{(s)\top}R_u^{(s)}e_i \tag{39}$$

for $h \in \mathcal{H}_g$, and zero for other heads. Squaring each valid entry and summing over those heads gives Equation (36). The proof requires no small-perturbation or softmax linearization. Fixed biases remain in the opposite-side activations. □

For the evaluated Llama architecture, the actual pre-softmax logits on valid positions are $Z/\sqrt{d_h}$. Their squared reconstruction loss is therefore $\mathcal{L}_R/d_h$; this common factor leaves the single-deletion ranking unchanged. It does not establish a full-mask optimum. It also makes no claim for additional weight-dependent nonlinear Q/K transformations.

B.3 Pairwise second-moment accumulation for QK-Wanda-R

RoPE acts within disjoint pairs of coordinates. For an even head width d_h, let $r = 1, \ldots, d_h/2$ index those pairs. In the evaluated Hugging Face Llama convention, pair r contains coordinates $\mathcal{P}_r = (r, r + d_h/2)$ in that order. Indexing a vector by $[\mathcal{P}_r]$ selects those two entries. Define the two-dimensional rotated key/query vectors $v_{g,r,u}^{(s)} = \widetilde{k}_{g,u}^{(s)}[\mathcal{P}_r]$ and $z_{h,r,t}^{(s)} = \widetilde{q}_{h,t}^{(s)}[\mathcal{P}_r]$. Their 2×2 second-moment matrices are

$$V_{K,r}^{(g,s)}(t) = \sum_{u\le t} v_{g,r,u}^{(s)}v_{g,r,u}^{(s)\top}, \tag{40}$$

$$V_{Q,r}^{(g,s)}(u) = \sum_{h\in\mathcal{H}_g}\sum_{t\ge u} z_{h,r,t}^{(s)}z_{h,r,t}^{(s)\top}. \tag{41}$$

These are sums of outer products, without subtracting a mean. Their off-diagonal entries retain products between the two coordinates. Let $R_{t,r}^{(s)}$ be the 2×2 part of the rotary map acting on pair r. If raw coordinate i is element $b \in \{1, 2\}$ of that pair, let $e_b \in \mathbb{R}^2$ select it. For a fixed sequence s, its rotated direction in the pair is $d_{t,b} = R_{t,r}^{(s)}e_b$. The query interaction is

$$\begin{aligned}\sum_{u\le t}(d_{t,b}^\top v_{g,r,u}^{(s)})^2 &= d_{t,b}^\top\left(\sum_{u\le t} v_{g,r,u}^{(s)}v_{g,r,u}^{(s)\top}\right)d_{t,b}\\ &= [R_{t,r}^{(s)\top}V_{K,r}^{(g,s)}(t)R_{t,r}^{(s)}]_{bb}.\end{aligned} \tag{42}$$

For keys, the same argument gives

$$\sum_{h\in\mathcal{H}_g}\sum_{t\ge u}(z_{h,r,t}^{(s)\top}R_{u,r}^{(s)}e_b)^2 = [R_{u,r}^{(s)\top}V_{Q,r}^{(g,s)}(u)R_{u,r}^{(s)}]_{bb}. \tag{43}$$

Multiplying by the input square at t or u, then averaging over the N sequences and multiplying by the deleted weight's w^2, recovers Equations (35) and (36).

To make the two-coordinate calculation explicit, consider one pair at one position. Write its rotation angle as θ, with $c_\theta = \cos\theta$ and $s_\theta = \sin\theta$. Its rotation matrix is

$$R = \begin{pmatrix} c_\theta & -s_\theta \\ s_\theta & c_\theta \end{pmatrix}. \tag{44}$$

Let V be either symmetric 2×2 second-moment matrix above, with entries $V_{11}, V_{12} = V_{21}, V_{22}$. Its two rotated diagonal entries are

$$[R^\top VR]_{11} = c_\theta^2V_{11} + 2c_\theta s_\theta V_{12} + s_\theta^2V_{22}, \tag{45}$$

$$[R^\top VR]_{22} = s_\theta^2V_{11} - 2c_\theta s_\theta V_{12} + c_\theta^2V_{22}. \tag{46}$$

These formulas follow by multiplying the two-dimensional vectors $(c_\theta, s_\theta)^\top$ and $(-s_\theta, c_\theta)^\top$ on both sides of V. The terms containing V_{12} explain why cross-coordinate products must be retained. Setting every rotary map to the identity matrix, which leaves each vector unchanged, recovers QK-Wanda.

C CAUSAL AND ROPE-AWARE ABLATION AT 8192 CALIBRATION TOKENS

We compared causal pre-RoPE QK-Wanda with causal-plus-RoPE QK-Wanda-R on Llama-3.2-1B/3B and Llama-3.1-8B at 50% and 80% Q/K sparsity. Both variants used separate whole-matrix budgets, one-shot masks, and identical calibration inputs for each seed. This protocol used 256 C4 sequences of length 32 and calibration seeds 0, 1, and 2; PPL evaluation used length 2048. The calibration budget and context differ from the main study, so these comparisons are not included in its win/loss counts.

Table 3: Causal pre-RoPE (C) and causal-plus-RoPE (R) with separate matrix budgets and 256×32 calibration tokens. PPL is the mean and sample standard deviation over three calibration seeds. Post-RoPE QK error is the pooled relative squared error (%), averaged over seeds. Bold marks lower displayed means within each checkpoint and sparsity, including ties.

Model	Sparsity	Score	WikiText-2 PPL	C4 PPL	Post-RoPE QK error (%)
Llama-3.2-1B	50%	C	**11.1933 ± 0.0196**	**17.8518 ± 0.0687**	0.20
		R	12.2233 ± 0.0195	21.0389 ± 0.0728	**0.18**
Llama-3.2-1B	80%	C	**192.6108 ± 0.4938**	**261.1174 ± 2.7507**	**2.88**
		R	232.3745 ± 0.8535	269.6802 ± 1.5493	3.04
Llama-3.2-3B	50%	C	**8.0829 ± 0.0012**	**11.8228 ± 0.0010**	0.12
		R	8.1256 ± 0.0014	11.8965 ± 0.0009	**0.11**
Llama-3.2-3B	80%	C	**40.9509 ± 0.4876**	**51.7862 ± 0.6025**	**1.23**
		R	44.5742 ± 0.4372	55.7155 ± 0.1331	1.25
Llama-3.1-8B	50%	C	**6.4453 ± 0.0002**	**9.8939 ± 0.0010**	0.12
		R	6.4663 ± 0.0008	9.9308 ± 0.0016	**0.11**
Llama-3.1-8B	80%	C	**14.0276 ± 0.0957**	**23.4658 ± 0.0695**	1.18
		R	15.3611 ± 0.0492	25.8150 ± 0.0377	**1.17**

The causal pre-RoPE variant had lower mean PPL on both datasets in all six model–sparsity pairs. The RoPE-aware variant had lower post-RoPE QK error in four of six pairs; its advantage was consistent at 50% but mixed at 80%. Thus, including RoPE in the reconstruction target did not improve PPL in this ablation. These results do not establish the same ordering for shared budgets or the main calibration protocol.

The held-out diagnostics used 16 C4 sequences of length 512, seed 20260810, starting from validation document row 1100. Post-RoPE QK and block-output errors use the pooled squared-error ratio in Equation (49). Causal pre-RoPE scoring also produced lower mean same-input block-output error in all six pairs and lower rollout block-output error in four of six. As with the main QK-error diagnostic, these stages need not rank masks in the same order as PPL. The unmasked variant is derived in Appendix B.1; this ablation compares only the two causal variants.

Variation across calibration seeds. For each of the two QK scoring variants, three checkpoints, and two sparsities, we computed the sample standard deviation of PPL over calibration seeds 0, 1, and 2. Across the six model–variant configurations at 50%, the standard deviations ranged from 0.00015 to 0.01964 on WikiText-2 and from 0.00089 to 0.07280 on C4. At 80%, the respective ranges were 0.04917–0.85345 and 0.03766–2.75072. These are within-configuration seed variations, not variation across models. The different calibration and allocation protocol prevents treating these ranges as uncertainty bounds for the main single-seed comparisons.

D EXPERIMENTAL DETAILS

D.1 CALIBRATION AND PERPLEXITY DATA

We sampled calibration text from the English C4 training shard `en/c4-train.00000-of-01024.json.gz`. Each base sample was a contiguous 2048-token window from one document. The main experiments and calibration-sensitivity studies used calibration seed 0; Appendix C reports a separate three-seed ablation. The main studies retained the same sampled windows across methods at a given sample count. The sample-count study used nested subsets of these windows. We selected the same 256-window base pool for both methods, giving 524,288 calibration tokens. For the context-length study, we split each selected window into contiguous segments of length T. Every tested length divided 2048, so segmentation retained exactly the same tokens, with no padding, packing, or attention across segments. We processed segments independently with the model's ordinary positional encoding. Each method used $524{,}288/T$ sequences at length T, with PPL evaluation fixed at length 2048. For the context-length study, we processed independent sequences in batches of $2048/T$, keeping 2048 tokens per calibration forward. Causal sums were computed within each sequence before aggregation over the batch, and positions restarted at each segment. The $T = 2048$ results were reused from the identical $N = 256$ sample-count configurations.

For Llama 2 and TinyLlama, we sampled 256 base windows with seed 0 using each checkpoint's native tokenizer, then segmented them into length-64 sequences. Within each checkpoint, every method and sparsity used the same token manifest. We did not reuse Llama 3 token IDs with these tokenizers.

WikiText-2 evaluation used the `wikitext-2-raw-v1` test split, joining text entries with two newline characters before tokenization. C4 evaluation used the first 1100 documents of `en/c4-validation.00000-of-00008.json.gz`, joined by spaces and truncated to at most 256×2048 tokens. We evaluated complete, non-overlapping 2048-token windows and discarded the final incomplete window. These same evaluation sets informed calibration and allocation choices on Llama-3.1-8B; no document range was reserved exclusively for development. The resulting token sequences were identical across pruning configurations for each checkpoint and tokenizer. We used batch size one, disabled the KV cache, and computed full-vocabulary next-token cross-entropy. To reduce memory use, we evaluated the vocabulary projection in 128-token chunks, preserving FP32 logits and the same cross-entropy calculation.

D.2 ZERO-SHOT EVALUATION

We evaluated the dense base checkpoints of TinyLlama-1.1B, Llama 2 7B/13B/70B, Llama-3.2-1B/3B, Llama-3.1-8B/70B, and Qwen2.5 0.5B/1.5B/3B/7B/14B/32B/72B, together with six Q/K-only pruned configurations per checkpoint: original row-wise WANDA and shared-budget QK-WANDA at 30%, 50%, and 80% sparsity. For TinyLlama and Llama 2 7B/13B, we also evaluated the six full-block configurations in Table 2. We applied the exact saved masks from the corresponding PPL experiments to the same pinned checkpoints, without repeating pruning or updating retained weights. In the Q/K-only setting, all non-Q/K parameters remained unchanged. Full-block masks additionally pruned V/O and MLP projections; embeddings, normalization parameters, biases, and the language-model head remained unchanged. We reused each checkpoint's dense reference across pruning scopes. Appendix M additionally evaluates nine matrix-budget Wanda configurations on the three largest checkpoints under this same task protocol.

We used LM Evaluation Harness v0.4.12 (EleutherAI, 2026), with no demonstrations or task subsampling, maximum context length 2048, batch size 4, and evaluation seed 1234. We used scaled dot-product attention throughout. The 0.5B–32B checkpoints used FP16 weights on A100 GPUs: Qwen2.5-32B used two GPUs and the smaller models used one. The Llama 70B checkpoints used FP16 weights on H200 GPUs, and Qwen2.5-72B used BF16 weights on two H200 GPUs under its validated numerical protocol. We evaluated the base models directly, without chat templates or an additional beginning-of-sequence token. The suite comprised BoolQ (Clark et al., 2019), RTE (Wang et al., 2018), HellaSwag (Zellers et al., 2019), WinoGrande (Sakaguchi et al., 2020), ARC-Easy and ARC-Challenge (Clark et al., 2018), and OpenBookQA (Mihaylov et al., 2018). We used the validation splits for BoolQ, RTE, HellaSwag, and WinoGrande (3270, 277, 10042, and 1267 examples), and the test splits for ARC-Easy, ARC-Challenge, and OpenBookQA (2376, 1172,

and 500 examples). Every configuration used identical evaluation examples and prompts. We used accuracy for BoolQ, RTE, and WinoGrande and length-normalized accuracy for HellaSwag, both ARC tasks, and OpenBookQA. The reported zero-shot score was their unweighted arithmetic mean, expressed as a percentage, computed before rounding individual task scores.

D.3 Reconstruction metrics

We sampled 16 held-out C4 validation sequences of length 512 with seed 20260810, restricting sampling to document rows at or after 1100. These documents were disjoint from the C4 perplexity subset and the calibration training split. Each checkpoint used its native tokenizer, with identical held-out tokens across methods and sparsities. We evaluated the saved masks from the PPL experiments, using calibration seed 0. A dense forward pass supplied the same reference inputs to the dense and pruned Q/K projections in every block.

Block-mean relative squared QK error. Let D be the number of transformer blocks, indexed by ℓ. Using the sequence and head notation from Appendix A.1, define the unscaled pre-RoPE product $B^{(s)}_{\ell,h} = Q^{(s)\top}_{\ell,h} K^{(s)}_{\ell,g(h)}$. The hatted product $\widehat{B}^{(s)}_{\ell,h}$ applies both saved Q/K masks on the same dense block input, retaining biases. The binary mask $C^{(s)}$ keeps causally allowed token pairs. The table column *Rel. QK error (%)* reports

$$E_{\mathrm{QK}} = \frac{100}{D} \sum_{\ell=1}^{D} \frac{\sum_{s,h} \left\| C^{(s)} \odot (\widehat{B}^{(s)}_{\ell,h} - B^{(s)}_{\ell,h}) \right\|_F^2}{\sum_{s,h} \left\| C^{(s)} \odot B^{(s)}_{\ell,h} \right\|_F^2}. \tag{47}$$

Within each block, we first summed squared errors and dense reference energies over sequences and query heads, including all query heads that share a key head under GQA, then divided. We averaged these relative errors equally across blocks, so blocks with larger QK magnitudes did not dominate the aggregate. All block denominators were positive. We took no square root: a value of 1 denotes mean relative *squared* error of 1%, rather than relative norm error of 1%. The common attention scale cancels within each block's ratio.

This diagnostic evaluates complete masks, including interactions among deletions, rather than summing individual deletion scores. It isolates each block's Q/K reconstruction on dense inputs; it does not include error propagated from previously pruned blocks. We retained FP16 projection outputs (BF16 for Qwen2.5-72B), computed QK dot products in FP32 with TF32 disabled, and accumulated squared-error sums in FP64. RoPE remained enabled in the dense forward pass that supplied block inputs; the measured products were taken before RoPE.

Cross-checkpoint aggregation. The abstract reports the mean of per-checkpoint relative reductions, not the relative reduction of mean errors. For checkpoint c at sparsity ρ, let $E^{\mathrm{QK}}_{c,\rho}$ and $E^{\mathrm{Wanda}}_{c,\rho}$ be the block-mean errors in Equation (47). We computed

$$\overline{\Delta}_{\rho} = \frac{100}{15} \sum_{c=1}^{15} \left(1 - \frac{E^{\mathrm{QK}}_{c,\rho}}{E^{\mathrm{Wanda}}_{c,\rho}} \right), \tag{48}$$

which gives 60.8% at 50% and 45.2% at 80%. All reference errors are positive at these sparsities. Section 4.3 discusses the relation to downstream metrics, and Appendix F.4 gives descriptive counts.

Additional reconstruction stages. For the earlier ablation in Appendix C, a reconstruction stage Y used the pooled relative squared error

$$E_Y = \frac{\sum_{s,\ell} \left\| \widehat{Y}^{(s)}_{\ell} - Y^{(s)}_{\ell} \right\|_F^2}{\sum_{s,\ell} \left\| Y^{(s)}_{\ell} \right\|_F^2}, \tag{49}$$

where hats denote pruned outputs. For post-RoPE QK logits, the sums also cover query heads and causally allowed token pairs. For full block outputs, the *same-input* setting supplies the dense reference input to each pruned block; the *rollout* setting instead supplies outputs from preceding pruned blocks. These stage metrics use $100E_Y$, without a square root, and are distinct from the equal-block average in Equation (47).

D.4 PARAMETER SCOPE OF Q/K-ONLY PRUNING

We counted unique model parameters from the pinned checkpoint configurations, instantiating model shapes without loading weight values and counting tied embeddings once. Q/K counts include only projection weights; biases remain dense. Let d be the hidden width, f the MLP intermediate width, d_h the head width, and H_Q, H_K the query/key head counts. For one block, the Q/K weights and all seven projection weights number

$$n_{QK} = d\, d_h(H_Q + H_K), \qquad n_{\text{linear}} = 2d\, d_h(H_Q + H_K) + 3df.$$

The model-level Q/K count is n_{QK} times the block count. The total-parameter denominator additionally includes embeddings, any untied language-model head, norms, and biases. At Q/K-only sparsity ρ, the removed fraction of total parameters is $\rho\, n_{QK,\text{total}}/n_{\text{total}}$, up to integer rounding. This measures zeroed weights, not storage or inference speedup.

Table 4: Scope of the Q/K-only experiments. B and M denote billions and millions of parameters. The last two columns use different denominators: the complete model and its seven block projection weight matrices.

Checkpoint	Total (B)	Q/K (M)	Q/K / total (%)	Q/K / block linear (%)
TinyLlama-1.1B	1.100	103.8	9.44	10.71
Llama-2-7B	6.738	1073.7	15.93	16.58
Llama-2-13B	13.016	2097.2	16.11	16.53
Llama-2-70B	68.977	6039.8	8.76	8.82
Llama-3.2-1B	1.236	83.9	6.79	8.62
Llama-3.2-3B	3.213	352.3	10.97	12.50
Llama-3.1-8B	8.030	671.1	8.36	9.62
Llama-3.1-70B	70.554	6039.8	8.56	8.82
Qwen2.5-0.5B	0.494	22.0	4.46	6.15
Qwen2.5-1.5B	1.544	77.1	4.99	5.88
Qwen2.5-3B	3.086	169.9	5.50	6.12
Qwen2.5-7B	7.616	411.0	5.40	6.30
Qwen2.5-14B	14.770	1509.9	10.22	11.43
Qwen2.5-32B	32.764	2013.3	6.14	6.45
Qwen2.5-72B	72.706	6039.8	8.31	8.60

D.5 IMPLEMENTATION AND REPRODUCIBILITY

The software environment was Python 3.10.14, PyTorch 2.4.1 with CUDA 12.1, Transformers 4.45.2, Accelerate 0.34.2, datasets 2.21.0, and LM Evaluation Harness 0.4.12. The primary server had four NVIDIA A100-PCIE-40GB GPUs. Of the 54 sparse Llama 70B PPL configurations, four retained A100 measurements: Wanda and separate-budget QK-Wanda at 50% sparsity on each checkpoint, using four GPUs per model. The other 50 configurations ran on H200 GPUs on Orix. The dense references came from the retained Wanda runs. Qwen2.5-72B PPL used H200 GPUs. The original Llama 70B table reconstruction measurements used one H200 per checkpoint; the full QK-error sparsity sweeps on the 70B/72B checkpoints used two H200 GPUs per model. Model weights were FP16 except for Qwen2.5-72B, which used native BF16 consistently across pruning methods and its dense reference to avoid nonfinite FP16 activations. The H200 jobs used Python 3.12.3. Pruning statistics were accumulated in FP32. Table 5 lists the pinned model/tokenizer revisions.

Table 5: Hugging Face repository identifiers and pinned revisions used for both model and tokenizer.

meta-llama/Llama-3.1-8B	
Revision	d04e592bb4f6aa9cfee91e2e20afa771667e1d4b
meta-llama/Llama-3.2-1B	
Revision	4e20de362430cd3b72f300e6b0f18e50e7166e08
meta-llama/Llama-3.2-3B	
Revision	13afe5124825b4f3751f836b40dafda64c1ed062
TinyLlama/TinyLlama-1.1B-intermediate-step-1431k-3T	
Revision	59f6f375b26bde864a6ca194a9a3044570490064
meta-llama/Llama-2-7b-hf	
Revision	01c7f73d771dfac7d292323805ebc428287df4f9
meta-llama/Llama-2-13b-hf	
Revision	5c31dfb671ce7cfe2d7bb7c04375e44c55e815b1
meta-llama/Llama-3.1-70B	
Revision	349b2ddb53ce8f2849a6c168a81980ab25258dac
meta-llama/Llama-2-70b-hf	
Revision	3aba440b59558f995867ba6e1f58f21d0336b5bb
Qwen/Qwen2.5-0.5B	
Revision	060db6499f32faf8b98477b0a26969ef7d8b9987
Qwen/Qwen2.5-1.5B	
Revision	8faed761d45a263340a0528343f099c05c9a4323
Qwen/Qwen2.5-3B	
Revision	3aab1f1954e9cc14eb9509a215f9e5ca08227a9b
Qwen/Qwen2.5-7B	
Revision	d149729398750b98c0af14eb82c78cfe92750796
Qwen/Qwen2.5-14B	
Revision	97e1e76335b7017d8f67c08a19d103c0504298c9
Qwen/Qwen2.5-32B	
Revision	1818d35814b8319459f4bd55ed1ac8709630f003
Qwen/Qwen2.5-72B	
Revision	efba10c8e54e91e0d9570ab5f7b51a958474d4cb

Integer counts and ties. Our row-wise Wanda implementation removed $\lceil \rho d \rceil$ weights per output row of width d; the matched row-wise QK-Wanda control used the same convention. Separate matrix budgets removed $\lceil \rho n_Q \rceil$ and $\lceil \rho n_K \rceil$ weights, and shared budgets removed $\lceil \rho(n_Q + n_K) \rceil$. This upward rounding is our implementation convention, not a property attributed to the original Wanda algorithm. The retained Llama-3.1-8B QK sweep used floor at the matrix/pair scope; its shared counts coincide with ceiling at the tested sparsities, while separate counts differ by at most one per projection. The 50% calibration studies also used floor, which equals ceiling for their even row widths and matrix sizes. The 70B/72B matrix-Wanda controls matched separate QK-Wanda counts exactly.

Scores were sorted in ascending order with a stable sort. Ties within a row followed increasing column index; matrix ties followed row-major order; shared ties used all Q entries followed by all K entries, each in row-major order. No independent per-projection or per-head score normalization was applied. Frozen source-manifest hashes, rather than the mutable working tree, identify executed code in the archived study plans (Table 6). Token manifests, exact deletion counts, masks, and metric files were retained.

Table 6: Frozen source-manifest SHA256 identifiers recorded in study plans; each hash continues on the following line. Retained runs also record mask and result hashes.

Study bundle	Source-manifest SHA256
qk_budget	91360064ecb2be26de3ffe3450262f6f 3f5961c4569c460920b14e3c1b6b788f
qk_budget_1b	10987a5ad0b19d3fbdea2dc1f83c8d6a 339c0460ab8e001d4e152e20d62fdc5b
qk_budget_3b	291a3cdcb13802ca1a7f9a8e5770eaf8 90d80f8de1d1b03a54e352da645d43a0
qk_budget_llama2	a0119c61deca3691885b1b8e1ca90556 6e762376e506277279de1a280a078a82
qk_budget_qwen25	400df4ddec5a87dbe4bd99c1c6511e03 957f361f4e07f3f6ececf1e5a3328d13
qk_budget_llama70b	5289268b6f2d339bd69bcf257e90a014 54a4c5ba65966e3f019d5cb3a06e2044
qk_budget_qwen72b	10079ee05d763a95fd651cda09416c57 08d945df5317c1d37f2b00abe6d2bcee

Artifact availability. The supplementary archive `qk_wanda_anonymous.zip` contains the pruning and evaluation implementation, tests, pinned checkpoint configurations, full-precision result bundles, and scripts for regenerating tables and figures. Its README gives environment setup and commands for repeating pruning and evaluation. Model weights, dataset text, and the large collection of original binary masks are not redistributed; masks can be regenerated, and their recorded hashes and allocation statistics are included. Infrastructure identifiers were anonymized without changing numerical measurements. The archive includes a file-integrity manifest and documents its validation and provenance conventions.

D.6 SHARED Q/K ALLOCATION ACROSS TRANSFORMER BLOCKS

We examined the shared-budget masks at 50% combined Q/K sparsity for all 15 checkpoints. Each run used calibration seed 0 and the causal pre-RoPE objective, with 8192 C4 training sequences of length 64 tokenized by the checkpoint's native tokenizer. Each block had its own shared Q/K budget: the fractions removed from Q and K could differ, while exactly 50% of their combined weights were removed. The profiles below use the same masks as the PPL and QK-error comparisons. Lines connect every measured block; markers are shown every fourth block, with different starting blocks for each model to reduce overlap.

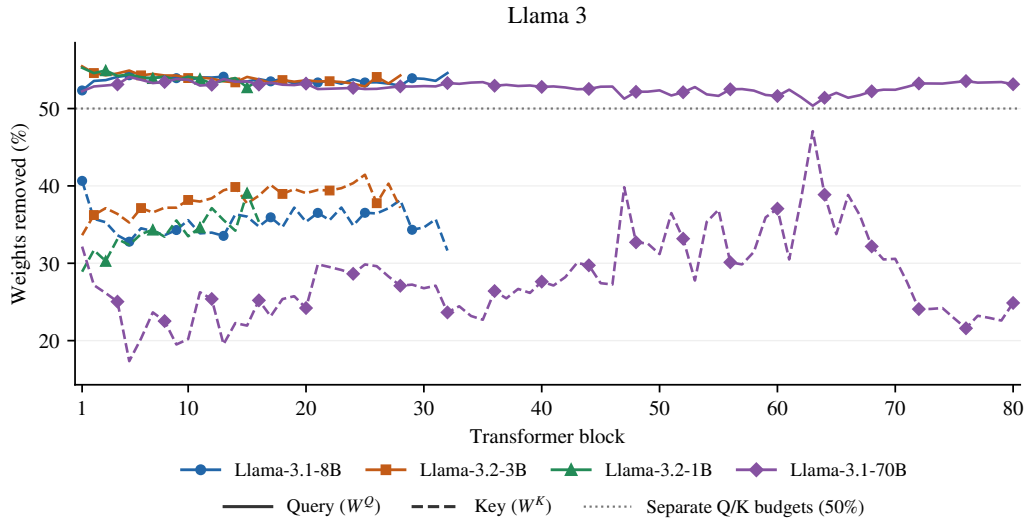

Figure 3: Shared 50% Q/K allocation for Llama-3.1-8B (blue), Llama-3.2-3B (orange), Llama-3.2-1B (green), and Llama-3.1-70B (purple). Color and marker shape identify the model; solid lines show W^Q and dashed lines show W^K. Each point is the percentage of weights removed within the projection. Actual block indices are retained: 1B, 3B, 8B, and 70B end at blocks 16, 28, 32, and 80, respectively. The dotted line marks separate 50% budgets. Percentages are measured within each projection; the shared budget is weighted by the Q/K parameter counts.

Llama 3: a consistent preference for retaining K weights. Every block in all four checkpoints removed a larger fraction of Q than K. Mean Q/K removal rates were 53.67%/35.33% for 8B, 53.92%/38.23% for 3B, 54.02%/33.93% for 1B, and 52.74%/28.09% for 70B. The 70B model therefore retained a larger fraction of K weights than the smaller models on average. Its allocation also varied more across blocks: Q/K standard deviations were 0.68/5.42 percentage points, compared with 0.42/1.69 on 8B, 0.57/1.71 on 3B, and 0.60/2.40 on 1B. On 70B, K removal ranged from 17.34% to 47.08%; none of the four profiles was strictly monotonic with depth.

Accounting for unequal projection sizes. The Q:K parameter-count ratios are 4:1 on 8B and 1B, 3:1 on 3B, and 8:1 on 70B. If Q contains r times as many weights as K, their removal fractions ρ_Q and ρ_K satisfy

$$\frac{r\rho_Q + \rho_K}{r+1} = 50\%.$$

Consequently, increasing Q sparsity by one percentage point requires decreasing K sparsity by r points. This constraint explains the larger amplitude of the K curves; it must be considered when comparing checkpoints. The allocations characterize the preference of the causal QK criterion under this calibration setup, rather than independently establishing downstream importance.

Llama 2 and TinyLlama. Figure 4 extends the same analysis to Llama 2 7B/13B/70B and TinyLlama 1.1B. The combined Q/K removal fraction was exactly 50% in every block, under the same calibration protocol described above.

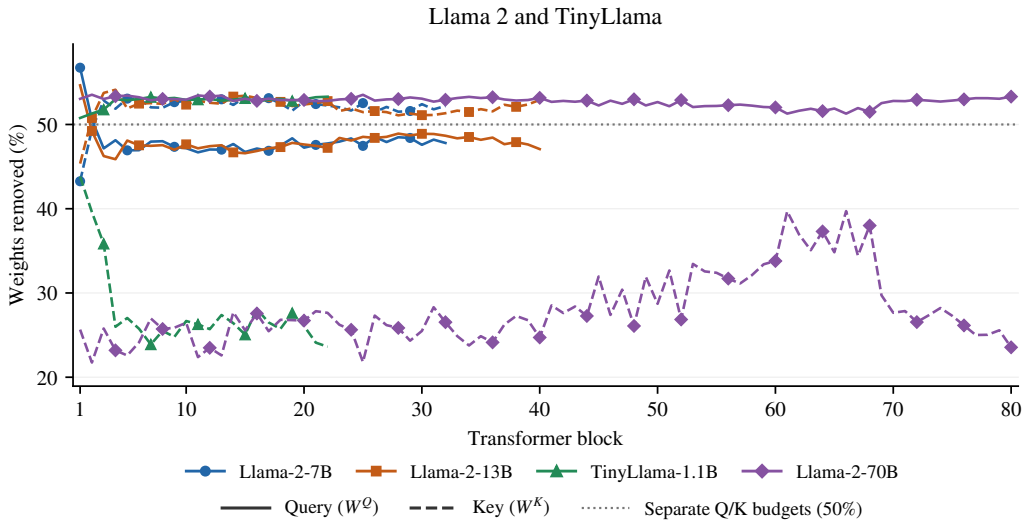

Figure 4: Shared 50% Q/K allocation for Llama-2-7B (blue), Llama-2-13B (orange), TinyLlama-1.1B (green), and Llama-2-70B (purple). Color and marker shape identify the model; solid lines show W^Q and dashed lines show W^K. The vertical axis is the fraction removed within each projection. TinyLlama, 7B, 13B, and 70B end at blocks 22, 32, 40, and 80, respectively. The dotted line marks separate 50% budgets. Percentages are measured within each projection; the shared budget is weighted by the Q/K parameter counts.

Similar allocations on Llama 2 7B and 13B. Mean Q/K removal rates were 47.98%/52.02% on 7B and 47.95%/52.05% on 13B. Both therefore retained a larger fraction of Q weights on average. K sparsity exceeded Q sparsity in 30 of 32 blocks on 7B and 39 of 40 on 13B; the exceptions were the first two blocks on 7B and the first block on 13B. Across-block standard deviations were 1.74 percentage points on 7B and 1.32 on 13B for either projection. These models have equally sized Q and K matrices, so their removal fractions average to 50% within each block.

Stronger K retention on TinyLlama and Llama 2 70B. Both checkpoints removed a larger fraction of Q than K in every block. Mean Q/K removal rates were 52.78%/27.80% on TinyLlama and 52.74%/28.04% on Llama 2 70B. Although these means were similar, their depth profiles differed: TinyLlama's K removal fell from 43.83% in the first block to 25.99% in the fourth, whereas 70B fluctuated between 21.74% and 39.72% over 80 blocks. The 70B Q/K standard deviations were 0.51/4.10 percentage points. Both have an 8:1 Q:K parameter-count ratio, so a modest Q sparsity increase above 50% permits a much larger reduction below 50% for K. The contrast with 7B/13B shows that allocation preferences vary within the Llama 2 family. It does not isolate an MHA-versus-GQA effect, because the checkpoints also differ in size and training.

Qwen2.5. We examined the shared-budget masks from all seven Qwen2.5 checkpoints in Section F.3, including 72B, at 50% combined Q/K sparsity. The curves report the fraction removed within each projection at the actual transformer block index. Because Q and K have unequal sizes under GQA, their combined sparsity is weighted by their parameter counts; it remained exactly 50% in every block.

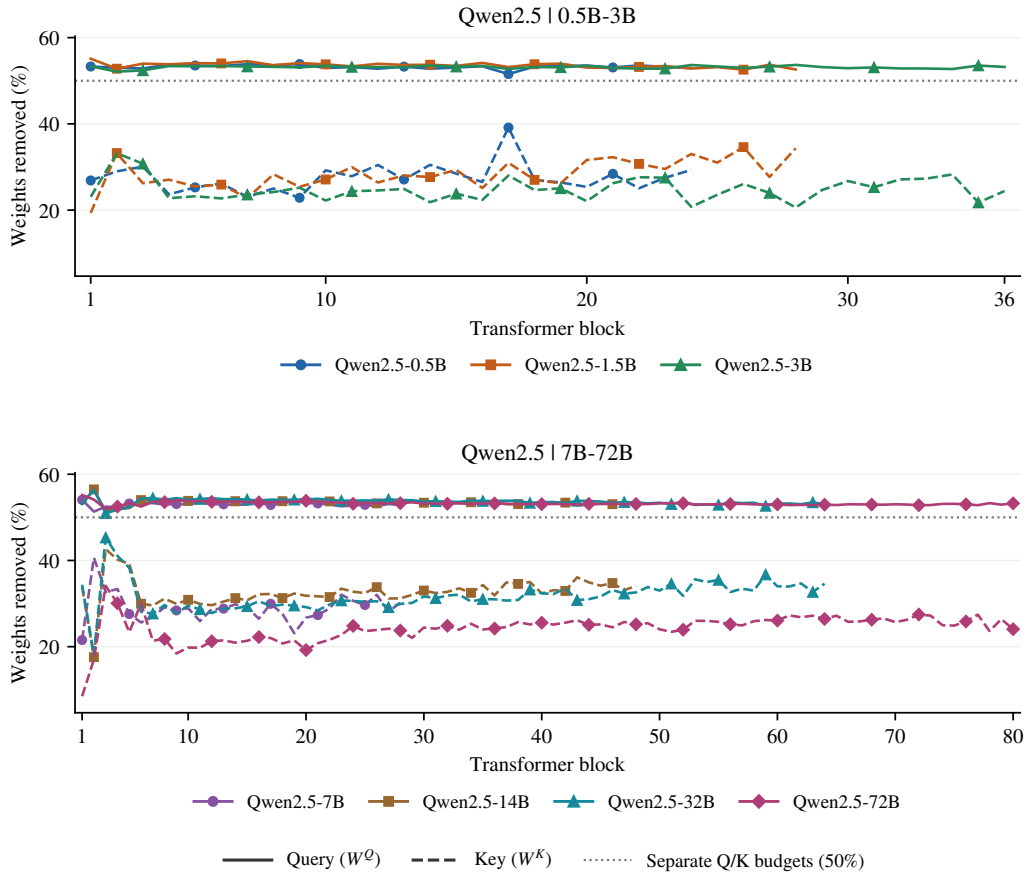

Figure 5: Shared 50% Q/K allocation on Qwen2.5, with 0.5B–3B models in the upper panel and 7B–72B models in the lower panel. Color and marker shape identify the checkpoint within each panel. Solid lines denote W^Q, dashed lines denote W^K, and the dotted line marks separate 50% budgets. Each curve ends at its checkpoint's final block; 72B extends to block 80. Percentages are measured within each projection; the shared budget is weighted by the Q/K parameter counts.

Observed allocation. Mean Q/K removal fractions were 53.21%/27.51% on Qwen2.5-0.5B, 53.59%/28.44% on Qwen2.5-1.5B, 53.14%/24.84% on Qwen2.5-3B, 53.01%/28.94% on Qwen2.5-7B, 53.47%/32.65% on Qwen2.5-14B, 53.65%/31.73% on Qwen2.5-32B, and 53.21%/24.29% on Qwen2.5-72B. Every block in all seven checkpoints removed a larger fraction of Q than K. The Q:K parameter-count ratios were 7:1, 6:1, 8:1, 7:1, 5:1, 5:1, and 8:1, respectively. On 72B, K removal ranged from 8.52% to 34.12% across 80 blocks, while Q removal ranged from 51.99% to 55.18%. The profiles describe the allocation favored by the causal QK criterion under this calibration setup; they do not by themselves establish the relative importance of Q and K for downstream performance.

E CALIBRATION SENSITIVITY

We varied calibration on Llama-3.1-8B at 50% Q/K sparsity with seed 0, comparing row-wise Wanda and separate-matrix QK-Wanda. The reported WikiText-2 test and C4 validation measurements informed selection; no separate development split was used. Evaluation context remained 2048.

Sample count. Table 7 varies the number of length-2048 windows. Both methods were weakly sensitive to sample count. We selected a common $N_* = 256$: all four PPL values were within 0.047% of their observed minima at $N = 2048$, which required eight times as many tokens. The WikiText-2/C4 improvements from 256 to 2048 windows were only 0.0020/0.0046 for Wanda and 0.0016/0.0003 for QK-Wanda.

Context length. Table 8 holds the 524,288-token pool fixed and varies segmentation. QK-Wanda attained its lowest PPL on both datasets at $T = 64$; $T = 128$ was within 0.0005. Wanda's minima occurred at 16 on WikiText-2 and 32 on C4; its length-64 values were only 0.0009/0.0033 higher. We selected $T = 64$ for both methods, giving 8192 independent sequences. These settings identify a stable operating region on the development checkpoint, not an optimal calibration for every model or sparsity.

Table 7: Sample-count sensitivity on Llama-3.1-8B at 50% Q/K sparsity, length 2048, and seed 0. Wanda uses row budgets and QK-Wanda separate matrix budgets. Evaluation context is 2048. Bold marks within-method, within-dataset minima at the displayed precision; † marks the selected common default, independently of boldface.

		Wanda PPL		QK-Wanda PPL	
Samples	Tokens	WikiText-2	C4	WikiText-2	C4
32	65,536	6.4165	9.8760	6.4585	9.9249
64	131,072	6.4163	9.8757	6.4573	9.9199
128	262,144	6.4161	9.8753	6.4583	9.9210
256†	524,288	6.4172	9.8781	6.4580	9.9186
512	1,048,576	6.4162	9.8759	6.4569	9.9194
1024	2,097,152	6.4162	9.8741	6.4565	9.9186
2048	4,194,304	**6.4152**	**9.8734**	**6.4564**	**9.9183**

Table 8: Context-length sensitivity on Llama-3.1-8B at 50% Q/K sparsity and seed 0, with exactly 524,288 calibration tokens. Wanda uses row budgets and QK-Wanda separate matrix budgets. Evaluation context is 2048. Bold marks within-method, within-dataset minima at the displayed precision; † marks the selected common default, independently of boldface.

		Wanda PPL		QK-Wanda PPL	
Length	Sequences	WikiText-2	C4	WikiText-2	C4
4	131,072	6.4076	9.8574	6.4722	9.9447
8	65,536	6.4036	9.8523	6.4599	9.9185
16	32,768	**6.4029**	9.8506	6.4522	9.9028
32	16,384	6.4043	**9.8494**	6.4458	9.8970
64†	8,192	6.4038	9.8527	**6.4437**	**9.8923**
128	4,096	6.4044	9.8563	6.4442	9.8925
256	2,048	6.4083	9.8608	6.4470	9.8979
512	1,024	6.4116	9.8669	6.4505	9.9048
1024	512	6.4157	9.8744	6.4535	9.9103
2048	256	6.4172	9.8781	6.4580	9.9186

F Additional Model Comparisons

We report the additional Q/K-only comparisons under the common protocol in Section 4. PPL and QK reconstruction entries use calibration seed 0; QK error follows Equation (47). The tables compare original row-wise WANDA with shared-budget QK-WANDA at 30%, 50%, and 80% sparsity; Appendix H also includes separate Q/K budgets and the full 0–90% sweeps. The additional matrix-budget Wanda control on 70B/72B checkpoints appears in Table 1 and Appendix M.

F.1 Llama 3

We compare row-wise Wanda with separate and shared QK-Wanda budgets under the common protocol in Section 4. The 8B separate-budget 50% point reuses the identical context-study configuration. Full curves appear in Appendix H; Table 9 collects the table sparsities.

On 8B, sharing the Q/K budget reduced PPL relative to separate budgets at every tested nonzero sparsity from 10% to 80% on both datasets, with negligible differences at 10–20%. At 90%, however, WikiText-2/C4 PPL increased from 40.35/62.22 to 63.48/109.86. At a shared 50% budget, mean removal rates were 53.67% for Q and 35.33% for K, with modest variation across blocks and no clear monotonic trend with depth. The criterion consistently preserved a larger fraction of K; Section D.6 compares the allocation by block for 1B, 3B, 8B, and 70B.

Against original WANDA, the shared allocation gave small PPL reductions on both datasets at 20–40%. At 50%, WikiText-2 PPL was nearly unchanged (an increase of 0.0012), while C4 PPL fell from 9.85 to 9.82. At 60–90%, WANDA outperformed both QK allocations; at 80%, its WikiText-2/C4 PPL was 8.89/13.54 versus 13.81/22.36 for shared QK-WANDA. Thus, sharing improved the QK criterion's allocation over most of the sweep, but did not resolve its degradation under aggressive pruning. Using the same reported WikiText-2 test/C4 validation sets for development, we selected shared Q/K budgets on 8B: they yielded lower PPL than separate budgets on both datasets at each of its three sparsities (30%, 50%, and 80%). We retained this allocation for 3B and 1B without retuning it.

On 3B, shared QK-WANDA again reduced PPL relative to WANDA on both datasets at 20–40% sparsity. At 40%, WikiText-2/C4 PPL fell from 7.92/11.53 to 7.90/11.49, reductions of 0.29%/0.38%. At 50%, WANDA was slightly better (8.06/11.79 versus 8.07/11.80), and it retained lower PPL than either QK allocation at every tested sparsity from 50% to 90%.

Sharing improved upon separate QK budgets on both datasets at 20–60%. At 70–80%, it lowered C4 PPL but increased WikiText-2 PPL: at 80%, separate and shared budgets gave 36.36/46.22 and 42.70/44.30, respectively, compared with 14.22/21.80 for WANDA. At 90%, sharing worsened both metrics. Thus, 3B and 8B both showed small gains over WANDA at moderate sparsity, whereas the benefit of sharing depended on model, sparsity, and dataset.

On 1B, sharing reduced both PPL values relative to separate QK budgets at 30–70% and 90% sparsity. The reductions were larger than on 3B at 50–70%: at 60%, separate and shared budgets gave WikiText-2/C4 PPL of 20.05/37.56 and 14.79/25.66, respectively. At 80%, sharing again improved C4 but worsened WikiText-2.

Compared with WANDA, shared QK-WANDA gave small reductions on both datasets at 10–30%, although the C4 differences at 10–20% were below 0.001 PPL. At 30%, WikiText-2/C4 PPL decreased from 9.83/14.15 to 9.81/14.14. At 40%, WikiText-2 was nearly tied, while C4 favored WANDA; at 50–80%, WANDA was better on both datasets. At 90%, shared QK-WANDA achieved lower PPL than both alternatives, with 335.84/323.81 versus 393.76/461.70 for WANDA. However, all three methods were severely degraded relative to the dense model (9.75/14.01). This reversal at extreme sparsity does not indicate preservation of dense-model quality.

Table 9 collects PPL, QK reconstruction error, and mean zero-shot accuracy for Llama-3.2-1B/3B and Llama-3.1-8B/70B. The shared allocation and calibration settings were retained across these checkpoints. Table 14 reports the individual task scores; Appendix F.1.1 discusses their relation to PPL.

Table 9: Q/K-only unstructured pruning across Llama 3 model sizes. Each model occupies four columns: WikiText-2 PPL, C4 PPL, mean zero-shot accuracy (%, seven tasks; Tables 14 and 16), and mean relative squared causal pre-RoPE QK error (%, Equation (47)). Pruning used calibration seed 0; QK-WANDA uses a shared Q/K budget within each block. Lower PPL and QK error and higher accuracy are better. Bold marks the better displayed result at each sparsity; displayed ties are bold in both rows.

		Llama-3.2-1B				**Llama-3.2-3B**			
Method	Sparsity	Wiki2 PPL	C4 PPL	Zero shot acc.	Rel. QK error (%)	Wiki2 PPL	C4 PPL	Zero shot acc.	Rel. QK error (%)
Dense	0%	9.75	14.01	54.14	0.00	7.81	11.33	61.74	0.00
WANDA	30%	9.83	14.15	**54.78**	0.23	7.85	11.41	61.35	0.14
QK-WANDA	30%	**9.81**	**14.14**	54.28	**0.08**	**7.84**	**11.38**	**61.55**	**0.06**
WANDA	50%	**10.31**	**15.03**	**54.22**	1.56	**8.06**	**11.79**	**61.11**	1.03
QK-WANDA	50%	10.65	16.24	53.51	**0.72**	8.07	11.80	60.99	**0.59**
WANDA	80%	**34.74**	**57.75**	**43.48**	15.49	**14.22**	**21.80**	**53.51**	**9.58**
QK-WANDA	80%	202.16	260.13	42.52	**14.84**	42.70	44.30	50.16	9.80

		Llama-3.1-8B				**Llama-3.1-70B**			
Method	Sparsity	Wiki2 PPL	C4 PPL	Zero shot acc.	Rel. QK error (%)	Wiki2 PPL	C4 PPL	Zero shot acc.	Rel. QK error (%)
Dense	0%	6.24	9.54	69.10	0.00	2.81	7.11	74.21	0.00
WANDA	30%	6.26	9.58	68.80	0.12	**2.82**	**7.12**	74.16	0.04
QK-WANDA	30%	**6.25**	**9.57**	**69.40**	**0.05**	**2.82**	**7.12**	**74.18**	**0.01**
WANDA	50%	**6.40**	9.85	68.43	0.79	**2.90**	**7.22**	73.60	0.37
QK-WANDA	50%	6.41	**9.82**	**69.39**	**0.37**	2.97	7.27	**73.79**	**0.16**
WANDA	80%	**8.89**	**13.54**	59.64	7.28	**4.98**	**9.89**	**66.64**	5.10
QK-WANDA	80%	13.81	22.36	**60.40**	**5.27**	18.98	32.37	66.45	**3.03**

The 70B task scores appear in Table 16.

F.1.1 ZERO-SHOT TASK PERFORMANCE

We evaluated the dense Llama-3.2-1B, Llama-3.2-3B, and Llama-3.1-8B checkpoints and the exact Q/K-only masks used in the PPL comparison. Table 14 reports the seven task scores and their unweighted mean under the protocol in Appendix D.2. These means also appear in Table 9. All other weights and biases remained unchanged.

The smaller checkpoints. On 1B, WANDA attained higher mean accuracy at all three sparsities, by 0.50, 0.71, and 0.96 percentage points at 30%, 50%, and 80%, respectively. The 30% result contrasts with the slightly lower PPL of QK-WANDA. On 3B, QK-WANDA improved mean accuracy by 0.21 points at 30%, but reduced it by 0.12 points at 50% and 3.35 points at 80%. At 80%, WANDA was better on five of seven tasks on 1B and six of seven on 3B; the lower mean accuracy of QK-WANDA agreed with its higher PPL on both checkpoints.

The 8B checkpoint. On 8B, QK-WANDA attained higher mean accuracy at every reported sparsity: 69.40% versus 68.80% at 30%, 69.39% versus 68.43% at 50%, and 60.40% versus 59.64% at 80%. The corresponding gains were 0.60, 0.96, and 0.76 percentage points. At 30%, it improved six tasks and tied on OpenBookQA; at 50%, it improved six tasks, with a small decrease on OpenBookQA.

The 80% comparison shows a disagreement between PPL and task accuracy. QK-WANDA had substantially higher WikiText-2/C4 PPL than WANDA (13.81/22.36 versus 8.89/13.54), yet higher accuracy on five of seven tasks and a higher mean. The largest task gains were on HellaSwag (4.82 points) and ARC-Challenge (2.13 points), while BoolQ and RTE decreased by 1.62 and 1.44 points. Both pruned models remained well below the dense mean of 69.10%. Thus, the PPL ordering did not

determine the task-accuracy ordering for these masks; neither metric alone captured all measured changes in predictive performance.

The 70B checkpoint. Table 16 reports all seven tasks. Shared QK-WANDA increased mean accuracy from 74.16% to 74.18% at 30% sparsity and from 73.60% to 73.79% at 50%. At 80%, WANDA was slightly better: 66.64% versus 66.45%, compared with the dense mean of 74.21%. The 50% result again separates the metrics: QK-WANDA had higher mean accuracy despite slightly higher PPL. At 80%, its lower causal pre-RoPE QK error (3.03% versus 5.10%) accompanied substantially higher WikiText-2/C4 PPL (18.98/32.37 versus 4.98/9.89), without a mean-accuracy advantage.

Interpretation. The comparison used one calibration seed and a fixed evaluation suite. Small mean gains, including the 8B results slightly above dense at 30%/50%, do not establish statistical significance or a general benefit from pruning. The higher 8B mean at 80% also does not imply improvement on every task or negate the measured increase in language-model loss. Together with the lower 1B/3B accuracies, these results show that the effect of Q/K-aware pruning depends on the checkpoint, sparsity, and downstream metric.

F.2 TINYLLAMA AND LLAMA 2

Table 10: Q/K-only unstructured pruning on Llama 2 and TinyLlama. Each model occupies four columns: WikiText-2 PPL, C4 PPL, mean zero-shot accuracy (%, seven tasks; Tables 14 and 16), and mean relative squared causal pre-RoPE QK error (%, Equation (47)). PPL and QK error use calibration seed 0; QK-WANDA uses a shared Q/K budget within each block. Lower PPL and QK error and higher accuracy are better. Bold marks the better displayed result at each sparsity, including ties.

		TinyLlama-1.1B (GQA)				**Llama-2-7B (MHA)**			
Method	Sparsity	Wiki2 PPL	C4 PPL	Zero shot acc.	Rel. QK error (%)	Wiki2 PPL	C4 PPL	Zero shot acc.	Rel. QK error (%)
Dense	0%	7.77	9.94	49.93	0.00	5.47	7.26	64.38	0.00
WANDA	30%	7.79	9.97	49.89	0.08	5.49	7.28	64.00	0.10
QK-WANDA	30%	**7.78**	**9.96**	**50.02**	**0.03**	**5.48**	**7.27**	**64.24**	**0.03**
WANDA	50%	7.93	10.12	**49.71**	0.44	5.62	7.42	63.43	0.72
QK-WANDA	50%	**7.85**	**10.04**	49.64	**0.17**	**5.53**	**7.36**	**63.92**	**0.29**
WANDA	80%	11.01	13.16	46.76	3.07	12.54	13.94	52.56	7.76
QK-WANDA	80%	**8.90**	**11.35**	**47.60**	**1.39**	**8.33**	**11.21**	**57.81**	**3.44**

		Llama-2-13B (MHA)				**Llama-2-70B (GQA)**			
Method	Sparsity	Wiki2 PPL	C4 PPL	Zero shot acc.	Rel. QK error (%)	Wiki2 PPL	C4 PPL	Zero shot acc.	Rel. QK error (%)
Dense	0%	4.88	6.73	67.05	0.00	3.32	5.71	71.50	0.00
WANDA	30%	4.90	6.74	66.77	0.06	**3.32**	**5.71**	71.50	0.04
QK-WANDA	30%	**4.89**	**6.73**	**66.92**	**0.02**	**3.32**	**5.71**	**71.62**	**0.01**
WANDA	50%	4.97	6.83	66.33	0.47	3.36	5.74	71.37	0.46
QK-WANDA	50%	**4.93**	**6.79**	**66.74**	**0.16**	**3.33**	**5.72**	**71.62**	**0.06**
WANDA	80%	8.06	10.20	56.35	6.04	4.55	6.91	65.85	7.57
QK-WANDA	80%	**6.27**	**8.58**	**64.09**	**2.15**	**3.63**	**5.97**	**71.76**	**1.06**

We evaluated the 3-trillion-token TinyLlama-1.1B base checkpoint (Zhang et al., 2024), using its native tokenizer and the same calibration and evaluation settings as the Llama 2 models. TinyLlama is independently trained and is not an official small Llama 2 checkpoint. It uses GQA with 32 query heads and 4 key/value heads, whereas Llama 2 7B/13B use MHA. The full comparison in Table 10 includes these checkpoints and Llama 2 70B; TinyLlama's full curves appear in Appendix H.3.

Shared-budget QK-WANDA achieved lower PPL than original WANDA on both datasets at every tested nonzero sparsity on TinyLlama. At 80%, WikiText-2/C4 PPL decreased from 11.01/13.16 to 8.90/11.35, relative reductions of 19.2%/13.7%. Sharing also improved on separate Q/K budgets: at 80%, separate budgets gave 9.71/12.34, and at 90%, sharing reduced PPL from 17.32/22.10 to 13.10/16.35. Unlike the near-overlapping allocations on Llama 2 7B/13B, these results show a substantial additional benefit from sharing at high sparsity. The improvement on this GQA checkpoint also shows that gains are not restricted to MHA models; differences in training and tokenization prevent attributing the cross-model pattern to attention architecture alone.

Llama 2 7B/13B/70B. All three tabulated sparsities improved both PPL metrics and mean zero-shot accuracy. At 80%, the 7B/13B WikiText-2 PPL decreased from 12.54/8.06 to 8.33/6.27; on 70B it decreased from 4.55 to 3.63. At 90%, 13B WikiText-2/C4 PPL decreased from 40.42/30.27 to 12.04/15.99, while 70B decreased from 11.71/13.92 to 4.75/7.01. Dense quality was not preserved at this sparsity. The full curves, including separate budgets, appear in Appendix H.

Task-level results on TinyLlama and Llama-2-7B Table 14 reports the individual task scores underlying the TinyLlama and 7B mean accuracies in Table 10, using the protocol in Appendix D.2. The same appendix table includes the 13B task scores. Each model section includes the dense checkpoint as a reference and compares the two pruning methods at matched sparsity.

At 80% sparsity, QK-WANDA improved five of seven task scores on TinyLlama and six of seven on 7B, raising mean accuracy by 0.84 and 5.24 percentage points, respectively. RTE favored WANDA on both models; OpenBookQA also favored it on TinyLlama. The dense means were 49.93% and 64.38%, so both methods still lost accuracy under this aggressive pruning setting.

At 30% and 50%, task-wise differences were mixed. In particular, TinyLlama's 50% QK-WANDA model improved five task scores but had a slightly lower mean (49.64% versus 49.71%), with decreases on RTE and ARC-Challenge. TinyLlama's 30% mean was marginally above dense (50.02% versus 49.93%). These small differences are observations for one pruning calibration seed, not evidence of a statistically significant gain or loss.

F.3 QWEN2.5

We extended the same comparison to six Qwen2.5 base checkpoints: 0.5B, 1.5B, 3B, 7B, 14B, and 32B (Qwen et al., 2024). All six use GQA, with respectively 14/2, 12/2, 16/2, 28/4, 40/8, and 40/8 query/key-value heads. These checkpoints include biases in their Q/K projections. We retained the biases in the projection outputs used to compute QK scores and left the bias parameters unchanged during pruning.

The experiments used the common protocol in Sections 4 and D, with native tokenizers, unchanged Q/K biases, and the same calibration tokens across methods within each checkpoint. The 0.5B–14B models used one A100 per run; 32B used two.

Table 11 compares original WANDA and shared-budget QK-WANDA at 30%, 50%, and 80% sparsity. We retained the shared allocation selected in the Llama 3 8B study without retuning it on Qwen. Appendix H.7 reports the full three-method curves for both datasets, and Appendix D.6 examines the per-block Q/K allocation at a shared 50% budget.

At 30% sparsity, shared-budget QK-WANDA achieved lower PPL than both original WANDA and separate Q/K budgets on both datasets for all six checkpoints. The reductions relative to WANDA were small: 0.07%–0.29% on WikiText-2 and 0.02%–0.13% on C4.

At 50% sparsity, shared-budget QK-WANDA also achieved lower PPL than both alternatives on both datasets for all six checkpoints. Relative to WANDA, the reductions ranged from 0.22% to 1.34% on WikiText-2 and from 0.09% to 0.80% on C4. Sharing had only a negligible effect over separate budgets on 32B: the WikiText-2 difference was below 0.001 PPL. At the shared 50% budget, every block removed a larger fraction of Q than K; mean removal rates across the six models ranged from 53.01% to 53.65% for Q and from 24.84% to 32.65% for K (Appendix D.6).

At 80% sparsity, sharing reduced PPL relative to separate Q/K budgets on both datasets for all six checkpoints. Compared with WANDA, shared QK-WANDA improved both metrics for 0.5B and 1.5B.

For 3B, 7B, and 14B, it lowered WikiText-2 PPL but increased C4 PPL. On 32B it increased both: 5.81/11.67 PPL on WikiText-2/C4 versus 5.61/10.82 for WANDA.

At 90% sparsity, outcomes varied across checkpoints and datasets. Separate-budget QK-WANDA had higher PPL than WANDA on both datasets for all six checkpoints. Sharing improved on WANDA on both datasets for 0.5B but increased PPL on both for 1.5B, 3B, 7B, and 14B. Results on 32B were mixed: shared QK-WANDA gave 7.98/15.61 PPL on WikiText-2/C4, compared with 7.97/15.73 for WANDA.

Table 11: Q/K-only unstructured pruning on Qwen2.5 base models. Each model occupies four columns: WikiText-2 PPL, C4 PPL, mean zero-shot accuracy (%), and mean relative squared causal pre-RoPE QK error (%, Equation (47)). PPL and QK error use calibration seed 0; QK-WANDA uses a shared Q/K budget within each block. Lower PPL and QK error and higher accuracy are better. Accuracy is the seven-task mean; Table 15 reports task-level scores. Bold compares methods at each sparsity before rounding, favoring lower PPL/QK error and higher accuracy.

		Qwen2.5-0.5B				**Qwen2.5-1.5B**				**Qwen2.5-3B**			
Method	Sparsity	Wiki2 PPL	C4 PPL	Zero shot acc.	Rel. QK error (%)	Wiki2 PPL	C4 PPL	Zero shot acc.	Rel. QK error (%)	Wiki2 PPL	C4 PPL	Zero shot acc.	Rel. QK error (%)
Dense	0%	13.07	20.38	50.86	0.00	9.26	15.12	61.64	0.00	8.03	13.35	65.29	0.00
WANDA	30%	13.12	20.43	50.48	0.07	9.30	15.15	61.45	0.08	8.06	**13.37**	65.22	0.06
QK-WANDA	30%	**13.09**	**20.40**	**50.97**	**0.02**	**9.27**	**15.13**	**61.69**	**0.02**	**8.05**	**13.37**	**65.31**	**0.02**
WANDA	50%	13.31	20.71	**50.69**	0.39	9.48	15.32	**61.57**	0.41	8.20	13.50	64.72	0.30
QK-WANDA	50%	**13.17**	**20.54**	50.57	**0.13**	**9.35**	**15.26**	60.58	**0.14**	**8.12**	**13.45**	**65.13**	**0.11**
WANDA	80%	16.24	25.07	**47.40**	2.23	12.38	18.88	55.71	2.62	10.44	**15.95**	**62.68**	1.96
QK-WANDA	80%	**15.07**	**23.07**	47.01	**1.15**	**11.50**	**18.59**	**57.06**	**1.23**	**10.06**	16.08	61.50	**0.97**

		Qwen2.5-7B				**Qwen2.5-14B**				**Qwen2.5-32B**			
Method	Sparsity	Wiki2 PPL	C4 PPL	Zero shot acc.	Rel. QK error (%)	Wiki2 PPL	C4 PPL	Zero shot acc.	Rel. QK error (%)	Wiki2 PPL	C4 PPL	Zero shot acc.	Rel. QK error (%)
Dense	0%	6.85	11.88	70.56	0.00	5.29	10.35	72.35	0.00	5.02	10.17	72.36	0.00
WANDA	30%	6.86	**11.89**	**70.45**	0.03	5.31	10.36	**72.62**	0.05	5.03	10.18	**72.48**	0.04
QK-WANDA	30%	**6.85**	**11.89**	70.44	**0.01**	**5.30**	**10.35**	72.59	**0.02**	**5.02**	**10.17**	72.32	**0.02**
WANDA	50%	6.90	11.93	**70.03**	0.14	5.41	10.43	72.84	0.28	5.07	10.21	**72.69**	0.24
QK-WANDA	50%	**6.89**	**11.92**	70.02	**0.05**	**5.35**	**10.40**	**72.95**	**0.12**	**5.04**	**10.20**	72.40	**0.10**
WANDA	80%	7.39	**12.53**	67.36	0.95	6.50	**11.53**	70.64	1.64	**5.61**	**10.82**	70.80	1.59
QK-WANDA	80%	**7.32**	12.62	**68.03**	**0.44**	**6.33**	11.79	**72.43**	**0.92**	5.81	11.67	**70.97**	**0.87**

Scaling to 72B. We evaluated Qwen2.5-72B (64 query heads and 8 key/value heads) under the same calibration and pruning protocol. All methods and the dense reference used BF16 weights because FP16 produced nonfinite activations even without pruning. Pruning and PPL evaluation used one H200 per configuration; QK-error evaluation used two.

Shared-budget QK-WANDA achieved lower PPL than both alternatives at 30%, 50%, 80%, and 90% sparsity on both datasets. At 80%, WikiText-2/C4 PPL decreased from 4.55/9.87 with WANDA to 4.24/9.52 (6.7%/3.5% reductions); at 90%, from 6.45/11.58 to 5.16/10.17. Unlike the 32B result at 80%, this checkpoint favored shared QK-WANDA on both datasets. Figures 30 and 31 report the full curves.

Shared-budget QK-WANDA reduced relative squared QK error at all three tabulated sparsities. At 50%, the error decreased from 0.18% with WANDA to 0.07%; at 80%, from 1.23% to 0.58%, accompanying lower PPL on both datasets. We used identical dense block inputs and the protocol in Equation (47), with BF16 projections, FP32 QK products, and FP64 reductions.

Table 12: Q/K-only unstructured pruning on Qwen2.5-72B. All methods use the same BF16 checkpoint, calibration tokens, and evaluation protocol. QK-WANDA uses shared Q/K budgets. Accuracy is the seven-task mean (Table 17). Lower PPL and QK error and higher zero-shot accuracy are better; bold compares the sparse methods at each sparsity before rounding.

Method	Sparsity	Wiki2 PPL	C4 PPL	Zero-shot acc. (%)	Rel. QK error (%)
Dense	0%	3.87	9.26	74.75	0.00
WANDA	30%	**3.88**	**9.26**	**74.68**	0.03
QK-WANDA	30%	**3.88**	**9.26**	74.61	**0.01**
WANDA	50%	3.92	9.28	**74.72**	0.18
QK-WANDA	50%	**3.90**	**9.27**	74.54	**0.07**
WANDA	80%	4.55	9.87	73.44	1.23
QK-WANDA	80%	**4.24**	**9.52**	**73.97**	**0.58**

F.3.1 ZERO-SHOT TASK PERFORMANCE

We evaluated all seven dense Qwen2.5 checkpoints and the exact saved Q/K-only masks used for the PPL and reconstruction-error comparisons. Each checkpoint had six pruned configurations: original row-wise WANDA and shared-budget QK-WANDA at 30%, 50%, and 80% sparsity. We used the seven full task splits and fixed evaluation protocol in Appendix D.2, without repeating pruning or updating retained weights. Tables 15 and 17 report individual scores; their unweighted means appear in Tables 11 and 12. The 72B evaluations used BF16 weights on two H200 GPUs.

The 0.5B and 1.5B checkpoints. At 30% sparsity, QK-WANDA increased mean accuracy by 0.49 and 0.24 percentage points, respectively. At 50%, it decreased the means by 0.13 and 0.99 points, despite reducing PPL on both datasets. At 80%, the mean decreased from 47.40% to 47.01% on 0.5B but increased from 55.71% to 57.06% on 1.5B. The 0.5B model improved on five tasks, but its 11.19-point decrease on RTE outweighed those gains in the mean. On 1.5B, five tasks improved, with the largest gain on BoolQ (9.79 points); ARC-Easy and WinoGrande decreased.

The 3B and 7B checkpoints. On 3B, QK-WANDA improved mean accuracy by 0.09 and 0.42 points at 30% and 50%, but decreased it by 1.18 points at 80% (62.68% to 61.50%). At 80%, the largest decrease was on RTE (8.66 points), whereas BoolQ and HellaSwag improved by 2.69 and 2.23 points. On 7B, means were almost unchanged at 30% and 50% (decreases of 0.02 and 0.01 points), and improved by 0.67 points at 80% (67.36% to 68.03%). Thus, the mixed WikiText-2/C4 PPL ordering at 80% accompanied lower mean accuracy on 3B and higher mean accuracy on 7B.

The 14B and 32B checkpoints. On 14B, the mean changed by -0.03, $+0.11$, and $+1.79$ points at 30%, 50%, and 80%, respectively. The 80% mean increased from 70.64% to 72.43%, the largest Qwen gain at that sparsity. Five tasks improved, led by BoolQ (8.35 points), while RTE and ARC-Easy decreased. On 32B, mean accuracy decreased by 0.17 and 0.29 points at 30% and 50%, but increased by 0.17 points at 80% (70.80% to 70.97%). The latter gain occurred despite higher PPL on both datasets, providing another example in which language-model loss and task accuracy ranked the masks differently.

The 72B checkpoint. At 30% and 50% sparsity, shared-budget QK-WANDA had slightly lower mean accuracy than WANDA: 74.61% versus 74.68%, and 74.54% versus 74.72%, respectively, despite lower PPL and QK error. At 80%, the ordering reversed: the mean increased from 73.44% to 73.97% (0.52 percentage points), while the dense reference scored 74.75%. Four tasks improved at 80% (BoolQ, HellaSwag, ARC-Easy, and ARC-Challenge), whereas RTE, WinoGrande, and OpenBookQA decreased.

The seven-task means and PPL can rank the same masks differently; Section 4.3 discusses this across families.

F.4 DESCRIPTIVE AGGREGATE COMPARISONS

All counts compare unrounded values, exclude dense points, and are descriptive rather than independent statistical trials. Lower PPL/error or higher accuracy determines the favorable direction. The 135 sweep configurations comprise 15 checkpoints and nine sparsities; task evaluation uses the 45 configurations at 30%, 50%, and 80%.

Table 13: Favorable comparisons / evaluated configurations. Shared QK-Wanda is compared with row-wise Wanda unless indicated. Accuracy uses the unweighted seven-task mean.

Subset / comparison	WikiText-2	C4	QK error	Accuracy
All sparsities	102/135	98/135	133/135	27/45
Tabulated sparsities	35/45	33/45	44/45	27/45
TinyLlama / Llama 2	35/36	35/36	36/36	11/12
Llama 3	13/36	13/36	34/36	6/12
Qwen2.5	54/63	50/63	63/63	10/21
30% sparsity	14/15	14/15	15/15	10/15
50% sparsity	11/15	12/15	15/15	7/15
80% sparsity	10/15	7/15	14/15	10/15
Shared vs. separate QK	111/135	117/135	121/135	—
Separate QK vs. row Wanda	94/135	85/135	124/135	—

Of the 135 shared-QK/row-Wanda comparisons, both PPL values decreased in 91, both increased in 26, and 18 had mixed signs. In 24 configurations, both PPL values increased despite lower QK error. Across the 315 individual task scores, 190 increased, 109 decreased, and 16 tied; tasks within a configuration are not independent replications.

G ZERO-SHOT TASK RESULTS

Table 14: Zero-shot accuracy (%) on TinyLlama, Llama 2, and Llama 3 with Q/K-only unstructured pruning. Each model section contains its dense reference and original row-wise WANDA versus shared-budget QK-WANDA. Task metrics follow Appendix D.2; Avg. is their unweighted mean. Bold marks the highest displayed accuracy within each model and sparsity, including ties.

Method	Sparsity	BoolQ	RTE	Hella Swag	Wino Grande	ARC-Easy	ARC-Challenge	OpenBook QA	Avg.
TinyLlama-1.1B									
Dense	0%	57.77	51.99	59.16	59.04	55.35	30.20	36.00	49.93
WANDA	30%	57.89	51.99	58.96	**58.56**	55.26	30.38	**36.20**	49.89
QK-WANDA	30%	**58.50**	**52.35**	**59.11**	58.41	**55.30**	**30.46**	36.00	**50.02**
WANDA	50%	56.94	**53.79**	58.24	57.70	55.47	**30.80**	35.00	**49.71**
QK-WANDA	50%	**57.68**	50.90	**58.76**	**58.41**	**55.56**	30.38	**35.80**	49.64
WANDA	80%	54.68	**55.23**	51.16	53.28	51.14	28.07	**33.80**	46.76
QK-WANDA	80%	**55.17**	52.35	**54.73**	**55.64**	**52.65**	**29.27**	33.40	**47.60**
Llama-2-7B									
Dense	0%	77.74	62.82	75.98	69.06	74.54	46.33	44.20	64.38
WANDA	30%	76.73	61.73	75.76	**68.75**	**74.79**	45.82	**44.40**	64.00
QK-WANDA	30%	**77.46**	**62.45**	**75.99**	68.67	74.75	**46.16**	44.20	**64.24**
WANDA	50%	75.96	**61.37**	74.80	68.43	74.45	44.80	44.20	63.43
QK-WANDA	50%	**76.57**	59.57	**75.99**	**68.90**	**75.17**	**46.42**	**44.80**	**63.92**
WANDA	80%	64.13	**54.87**	60.34	56.12	61.41	33.28	37.80	52.56
QK-WANDA	80%	**72.48**	53.79	**69.23**	**63.61**	**66.33**	**37.80**	**41.40**	**57.81**
Llama-2-13B									
Dense	0%	80.61	65.34	79.38	72.14	77.53	49.15	45.20	67.05
WANDA	30%	80.37	63.18	79.33	71.98	77.48	**49.49**	**45.60**	66.77
QK-WANDA	30%	**80.73**	**63.90**	**79.42**	**72.22**	**77.57**	48.98	**45.60**	**66.92**
WANDA	50%	80.12	63.18	78.69	**71.43**	**77.40**	48.46	45.00	66.33
QK-WANDA	50%	**80.98**	**63.90**	**79.39**	71.03	76.98	**48.89**	**46.00**	**66.74**
WANDA	80%	67.00	53.43	67.60	57.93	67.55	39.76	41.20	56.35
QK-WANDA	80%	**79.17**	**62.82**	**75.35**	**71.35**	**71.59**	**43.52**	**44.80**	**64.09**
Llama-3.2-1B									
Dense	0%	64.01	56.68	63.67	60.62	60.44	36.35	37.20	54.14
WANDA	30%	**64.40**	**60.29**	63.49	**61.09**	60.31	**36.86**	37.00	**54.78**
QK-WANDA	30%	64.37	57.40	**63.65**	60.30	**60.40**	36.43	**37.40**	54.28
WANDA	50%	63.09	**63.54**	61.78	**59.91**	**59.22**	**35.84**	36.20	**54.22**
QK-WANDA	50%	**63.85**	57.76	**62.78**	59.35	58.67	35.58	**36.60**	53.51
WANDA	80%	59.60	**51.99**	**41.29**	**53.12**	**45.37**	**25.43**	27.60	**43.48**
QK-WANDA	80%	**61.07**	51.62	37.29	51.22	41.71	25.34	**29.40**	42.52
Llama-3.2-3B									
Dense	0%	73.24	54.51	73.59	70.01	71.72	46.08	43.00	61.74
WANDA	30%	**73.49**	53.43	73.40	69.22	**71.63**	45.65	**42.60**	61.35
QK-WANDA	30%	73.21	**54.15**	**73.56**	**69.30**	**71.63**	**46.42**	**42.60**	**61.55**
WANDA	50%	72.72	**57.04**	72.25	**69.06**	69.53	44.80	**42.40**	**61.11**
QK-WANDA	50%	**72.78**	54.87	**73.31**	68.59	**70.16**	**45.82**	41.40	60.99
WANDA	80%	63.00	**60.29**	**60.12**	**58.88**	**57.83**	**37.29**	**37.20**	**53.51**
QK-WANDA	80%	**65.66**	48.01	58.25	57.30	53.41	33.11	35.40	50.16
Llama-3.1-8B									
Dense	0%	82.11	69.68	78.88	73.64	81.10	53.50	44.80	69.10
WANDA	30%	81.80	68.23	78.60	73.09	81.10	53.58	**45.20**	68.80
QK-WANDA	30%	**82.32**	**70.76**	**79.05**	**73.64**	**81.19**	**53.67**	**45.20**	**69.40**
WANDA	50%	81.19	69.68	77.76	72.06	80.09	53.41	**44.80**	68.43
QK-WANDA	50%	**82.54**	**71.84**	**79.13**	**73.48**	**80.56**	**53.58**	44.60	**69.39**
WANDA	80%	**68.26**	**61.37**	68.65	66.14	70.24	41.81	41.00	59.64
QK-WANDA	80%	66.64	59.93	**73.47**	**66.22**	**70.37**	**43.94**	**42.20**	**60.40**

Table 15: Zero-shot accuracy (%) on Qwen2.5 base checkpoints with Q/K-only unstructured pruning. Each model section contains its dense reference and original row-wise WANDA versus shared-budget QK-WANDA. Task metrics follow Appendix D.2; Avg. is their unweighted mean. Bold marks the highest displayed accuracy within each model and sparsity, including ties.

Method	Sparsity	BoolQ	RTE	Hella Swag	Wino Grande	ARC-Easy	ARC-Challenge	OpenBook QA	Avg.
Qwen2.5-0.5B									
Dense	0%	62.42	58.84	52.15	56.27	58.71	32.42	35.20	50.86
WANDA	30%	61.10	58.48	51.95	56.04	58.29	**32.51**	35.00	50.48
QK-WANDA	30%	**61.71**	**59.21**	**52.16**	**57.46**	**58.50**	32.34	**35.40**	**50.97**
WANDA	50%	60.86	56.32	51.63	**56.99**	**61.36**	**33.11**	34.60	**50.69**
QK-WANDA	50%	**62.51**	**57.76**	**52.00**	55.56	59.13	32.00	**35.00**	50.57
WANDA	80%	56.54	**60.29**	47.49	**53.67**	54.17	28.24	31.40	**47.40**
QK-WANDA	80%	**60.24**	49.10	**49.10**	53.59	**54.67**	**29.95**	**32.40**	47.01
Qwen2.5-1.5B									
Dense	0%	72.94	70.04	67.77	63.30	71.51	45.14	40.80	61.64
WANDA	30%	**73.09**	68.95	**67.82**	63.54	**71.13**	44.62	**41.00**	61.45
QK-WANDA	30%	**73.09**	**70.76**	67.75	**63.85**	70.62	**44.97**	40.80	**61.69**
WANDA	50%	**73.27**	**69.68**	67.07	62.27	**73.32**	**44.80**	40.60	**61.57**
QK-WANDA	50%	71.50	67.15	**67.71**	**63.54**	69.44	43.94	**40.80**	60.58
WANDA	80%	58.81	60.29	61.47	**61.25**	**70.54**	39.42	38.20	55.71
QK-WANDA	80%	**68.59**	**61.01**	**63.19**	60.93	67.26	**39.85**	**38.60**	**57.06**
Qwen2.5-3B									
Dense	0%	77.22	75.45	73.53	68.35	73.11	47.35	42.00	65.29
WANDA	30%	77.13	73.65	73.46	**69.38**	72.90	**47.61**	**42.40**	65.22
QK-WANDA	30%	**77.58**	**74.01**	**73.60**	68.59	**73.82**	47.35	42.20	**65.31**
WANDA	50%	77.06	**72.20**	72.88	**68.59**	73.15	46.93	42.20	64.72
QK-WANDA	50%	**77.95**	71.48	**73.28**	**68.59**	**74.71**	**47.53**	**42.40**	**65.13**
WANDA	80%	71.04	**77.26**	68.01	**65.43**	**73.06**	43.77	**40.20**	**62.68**
QK-WANDA	80%	**73.73**	68.59	**70.24**	64.01	69.53	**44.20**	**40.20**	61.50
Qwen2.5-7B									
Dense	0%	84.65	81.59	78.94	73.09	77.36	51.11	47.20	70.56
WANDA	30%	**84.53**	**81.95**	78.90	**73.16**	77.27	**50.77**	46.60	**70.45**
QK-WANDA	30%	84.46	81.59	**78.99**	72.93	**77.31**	**50.77**	**47.00**	70.44
WANDA	50%	**84.74**	**80.87**	78.85	**71.59**	76.68	50.51	**47.00**	**70.03**
QK-WANDA	50%	84.62	80.51	**79.07**	71.51	**76.98**	**51.28**	46.20	70.02
WANDA	80%	**82.94**	76.17	77.66	**69.46**	72.47	47.18	**45.60**	67.36
QK-WANDA	80%	82.91	**77.98**	**78.67**	68.90	**73.27**	**49.06**	45.40	**68.03**
Qwen2.5-14B									
Dense	0%	85.26	80.14	82.88	74.98	79.12	58.87	45.20	72.35
WANDA	30%	85.41	**80.14**	82.88	**76.40**	79.21	58.70	**45.60**	**72.62**
QK-WANDA	30%	**85.54**	**80.14**	**82.96**	75.06	**79.59**	**59.47**	45.40	72.59
WANDA	50%	84.19	**81.59**	82.82	75.69	80.68	58.70	**46.20**	72.84
QK-WANDA	50%	**85.93**	79.42	**82.92**	**76.09**	**80.89**	**59.22**	**46.20**	**72.95**
WANDA	80%	77.49	**80.87**	79.94	72.93	**82.20**	55.03	46.00	70.64
QK-WANDA	80%	**85.84**	78.34	**81.71**	**74.19**	81.48	**57.42**	**48.00**	**72.43**
Qwen2.5-32B									
Dense	0%	87.16	81.59	84.07	75.30	78.28	55.72	44.40	72.36
WANDA	30%	87.31	**81.23**	**84.06**	**75.61**	**78.28**	**56.48**	**44.40**	**72.48**
QK-WANDA	30%	**87.40**	80.87	84.02	75.30	78.20	56.23	44.20	72.32
WANDA	50%	87.13	**82.31**	83.80	**75.85**	**78.07**	**56.48**	45.20	**72.69**
QK-WANDA	50%	**87.19**	80.14	**83.89**	**75.85**	77.69	56.23	**45.80**	72.40
WANDA	80%	84.19	80.51	81.75	73.95	**75.55**	**54.27**	45.40	70.80
QK-WANDA	80%	**84.68**	**80.87**	**82.75**	**74.66**	74.33	53.92	**45.60**	**70.97**

Table 16: Zero-shot accuracy (%) on Llama 2 and Llama 3.1 70B with Q/K-only unstructured pruning. Each model section contains its dense reference and original row-wise WANDA versus shared-budget QK-WANDA. Task metrics follow Appendix D.2; Avg. is their unweighted mean. Bold marks the highest displayed accuracy within each model and sparsity, including ties.

Method	Sparsity	BoolQ	RTE	Hella Swag	Wino Grande	ARC-Easy	ARC-Challenge	OpenBook QA	Avg.
Llama-2-70B									
Dense	0%	83.76	67.87	83.81	77.98	80.98	57.34	48.80	71.50
WANDA	30%	83.43	**68.95**	83.72	77.82	80.72	**57.42**	48.40	71.50
QK-WANDA	30%	**83.64**	**68.95**	**83.79**	**77.90**	**81.02**	**57.42**	**48.60**	**71.62**
WANDA	50%	82.63	**69.31**	83.38	**77.98**	80.89	57.00	48.40	71.37
QK-WANDA	50%	**83.94**	68.95	**83.79**	77.90	**81.02**	**57.17**	**48.60**	**71.62**
WANDA	80%	76.51	65.70	76.66	68.03	78.11	50.34	45.60	65.85
QK-WANDA	80%	**84.04**	**70.40**	**82.54**	**77.27**	**81.44**	**56.66**	**50.00**	**71.76**
Llama-3.1-70B									
Dense	0%	85.44	69.68	85.03	79.95	86.66	64.93	47.80	74.21
WANDA	30%	**85.35**	70.04	**84.86**	79.56	86.36	**64.76**	**48.20**	74.16
QK-WANDA	30%	85.26	**70.76**	84.85	**79.64**	**86.62**	64.33	47.80	**74.18**
WANDA	50%	84.74	**70.04**	84.32	79.01	85.94	62.97	**48.20**	73.60
QK-WANDA	50%	**85.11**	**70.04**	**84.54**	**79.24**	**86.20**	**63.82**	47.60	**73.79**
WANDA	80%	76.06	68.59	**76.62**	**69.14**	80.89	54.01	**41.20**	**66.64**
QK-WANDA	80%	**77.80**	**71.12**	73.85	64.88	**81.61**	**55.72**	40.20	66.45

Table 17: Zero-shot accuracy (%) on Qwen2.5-72B with Q/K-only unstructured pruning. Dense is the unpruned reference; QK-WANDA uses shared Q/K budgets. We report accuracy for BoolQ, RTE, and WinoGrande, and length-normalized accuracy for HellaSwag, ARC-Easy/Challenge, and OpenBookQA. Avg. is the unweighted mean of the seven unrounded scores. Higher is better; bold marks the best displayed result at each sparsity, including ties.

Method	Sparsity	BoolQ	RTE	Hella Swag	Wino Grande	ARC-Easy	ARC-Challenge	OpenBook QA	Avg.
Dense	0%	89.17	77.26	86.06	78.14	83.38	62.63	46.60	74.75
WANDA	30%	**89.27**	**77.26**	**86.13**	77.51	**83.33**	**62.46**	**46.80**	**74.68**
QK-WANDA	30%	89.17	**77.26**	86.04	**77.58**	83.16	**62.46**	46.60	74.61
WANDA	50%	89.11	**76.90**	85.90	**77.82**	83.42	**62.12**	**47.80**	**74.72**
QK-WANDA	50%	**89.42**	76.53	**85.96**	77.58	**83.84**	61.86	46.60	74.54
WANDA	80%	87.55	**77.26**	83.90	**76.40**	83.25	60.15	**45.60**	73.44
QK-WANDA	80%	**89.27**	75.09	**85.06**	76.24	**85.23**	**62.29**	44.60	**73.97**

H PERPLEXITY ACROSS SPARSITY LEVELS

Each checkpoint has WikiText-2 and C4 curves, split into 0–50% and 50–90% panels with independent logarithmic ranges. All use the protocol in Section 4; dense points are shared references. Besides the three default curves, Llama 2 7B includes RoPE and row-budget controls; the largest checkpoints include matrix-budget Wanda (Appendix M).

H.1 LLAMA 2 13B

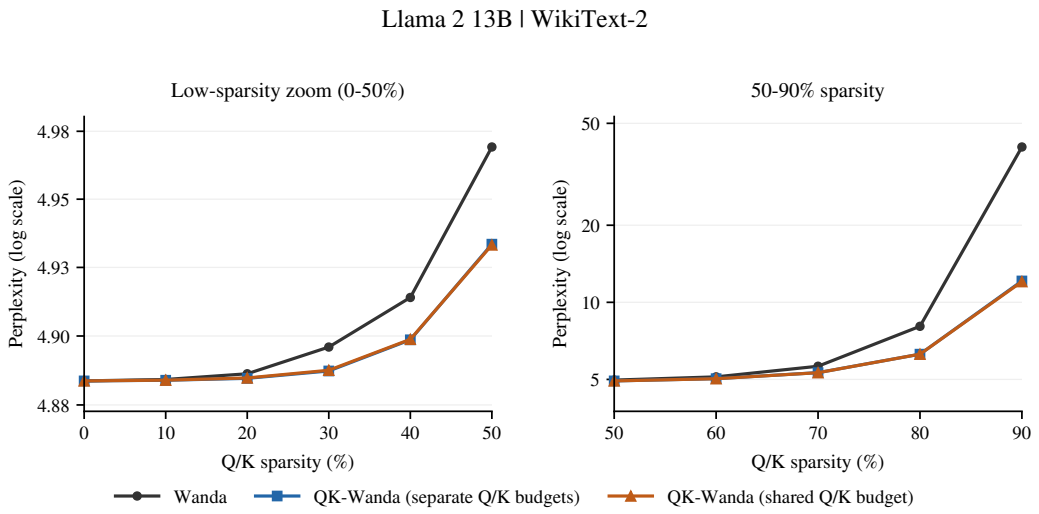

Figure 6: Llama-2-13B on WikiText-2. Original row-wise Wanda and separate/shared QK-Wanda budgets are compared over 0–50% and 50–90% sparsity with independent logarithmic PPL ranges.

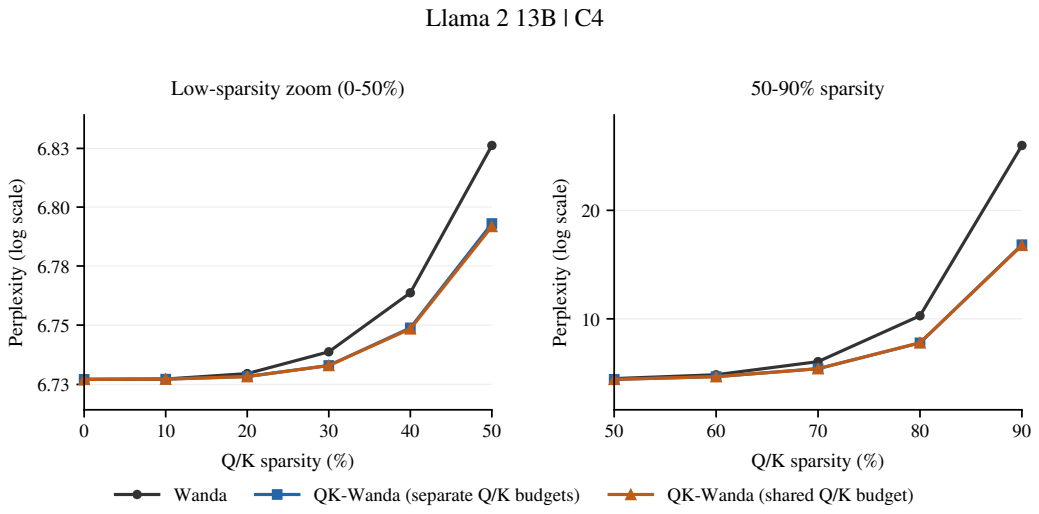

Figure 7: Llama-2-13B on C4, under the same pruning and calibration protocol as Figure 6.

H.2 LLAMA 2 7B

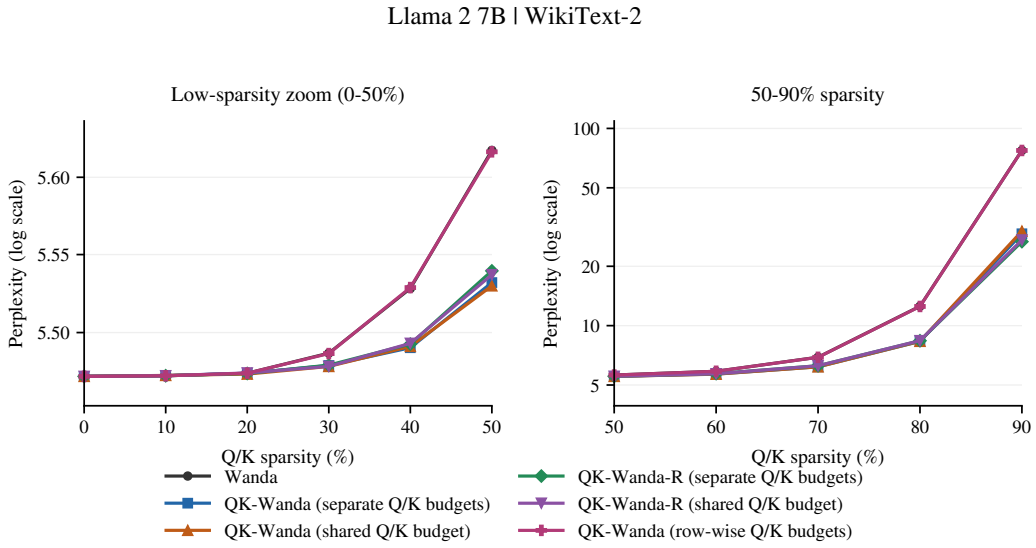

Figure 8: Llama-2-7B on WikiText-2. Six curves compare original row-wise WANDA, causal pre-RoPE QK-WANDA with row-wise, separate matrix, or shared Q/K budgets, and causal RoPE-aware QK-WANDA-R with separate or shared budgets. The panels use independent logarithmic PPL ranges. Table 18 reports reconstruction and zero-shot results.

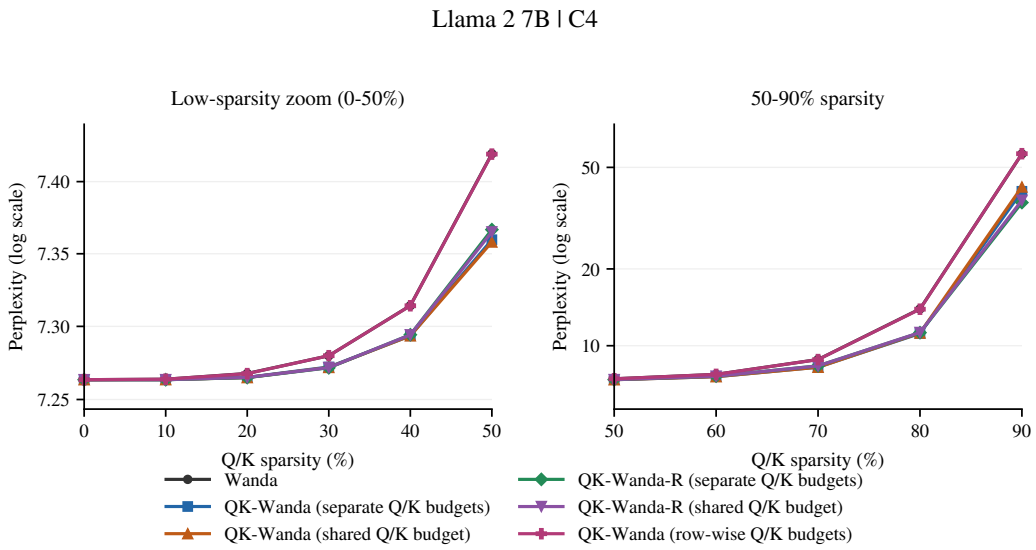

Figure 9: Llama-2-7B on C4 with the same six methods and matched calibration as Figure 8. Every QK-Wanda variant includes the causal mask; QK-Wanda-R additionally incorporates RoPE in its scoring loss. The panels have independent logarithmic PPL ranges.

Scoring and allocation ablation on Llama-2-7B. We evaluated causal pre-RoPE scores with row-wise, separate matrix, and shared Q/K budgets, and the RoPE-aware scores from Appendix B with separate and shared budgets. All runs used the same checkpoint, frozen calibration tokens, context lengths, seed, FP16 precision, and ceiling rule for deletion counts. At sparsity s, row-wise QK-Wanda removed exactly $\lceil 4096s \rceil$ weights from every Q and K row, matching Wanda's integer quotas. This control fixes allocation granularity while changing the score; causal scores need not preserve within-row rankings (Appendix A.3). Row-wise pruning and its PPL and zero-shot evaluations ran on NVIDIA H200 GPUs. The existing Wanda, separate/shared causal, and RoPE pruning sweeps ran on A100 GPUs.

At 30%, 50%, and 80%, we replayed the saved masks for the seven zero-shot tasks in Appendix D.2, including the original separate-budget causal masks. QK error used the same causal *pre-RoPE* diagnostic on common dense inputs for every method (Equation (47)); it does not measure the RoPE-aware objective itself. These errors were evaluated on the original A100 hardware, with a repeated shared-budget control reproducing the published block-level errors. Table 18 includes all six methods and the dense reference.

Table 18: Scoring and budget ablation on Llama-2-7B with Q/K-only unstructured pruning. QK-Wanda uses causal pre-RoPE scores with row-wise, separate matrix, or shared Q/K budgets. QK-Wanda-R additionally includes RoPE, with separate or shared budgets. Wanda uses row-wise budgets. All methods share the pre-RoPE error diagnostic (Equation (47)). Bold selects the best displayed result among six methods at each sparsity, including ties.

Method	Sparsity	WikiText-2 PPL	C4 PPL	Avg. acc. (%)	Rel. QK error (%)
Dense	0%	5.47	7.26	64.38	0.00
WANDA	30%	5.49	7.28	64.00	0.10
QK-WANDA (row-wise)	30%	5.49	7.28	63.95	0.10
QK-WANDA (separate)	30%	**5.48**	**7.27**	**64.32**	**0.03**
QK-WANDA (shared)	30%	**5.48**	**7.27**	64.24	**0.03**
QK-WANDA-R (separate)	30%	**5.48**	**7.27**	64.15	0.04
QK-WANDA-R (shared)	30%	**5.48**	**7.27**	64.21	0.04
WANDA	50%	5.62	7.42	63.43	0.72
QK-WANDA (row-wise)	50%	5.62	7.42	63.42	0.72
QK-WANDA (separate)	50%	**5.53**	**7.36**	63.91	**0.28**
QK-WANDA (shared)	50%	**5.53**	**7.36**	63.92	0.29
QK-WANDA-R (separate)	50%	5.54	7.37	63.98	0.34
QK-WANDA-R (shared)	50%	5.54	7.37	**64.02**	0.35
WANDA	80%	12.54	13.94	52.56	7.76
QK-WANDA (row-wise)	80%	12.51	13.90	52.51	7.78
QK-WANDA (separate)	80%	**8.32**	**11.17**	58.03	**3.40**
QK-WANDA (shared)	80%	8.33	11.21	57.81	3.44
QK-WANDA-R (separate)	80%	8.38	11.25	**58.13**	4.29
QK-WANDA-R (shared)	80%	8.39	11.27	57.82	4.33

Effect of allocation. Row-wise causal QK-Wanda closely tracked Wanda: absolute PPL differences were below 0.001 on both datasets through 60% sparsity. At 80%, its WikiText-2/C4 PPL was 12.51/13.90, compared with Wanda's 12.54/13.94; at 90%, the differences remained below 0.10 and had opposite signs across datasets. Its mean zero-shot accuracy was lower by 0.05, 0.01, and 0.05 percentage points at 30%, 50%, and 80%, and its QK error was slightly higher at all three levels before rounding. Given the single seed and different GPU platforms for PPL and task evaluation, these small changes do not establish a reliable advantage from changing the score under fixed row quotas.

Allowing the causal QK scores to redistribute deletions across rows had a much larger effect on this checkpoint. At 80%, separate and shared budgets achieved WikiText-2/C4 PPL of 8.32/11.17 and 8.33/11.21, respectively, and mean accuracy of 58.03% and 57.81%, versus 52.51% for row-wise QK-Wanda. Separate budgets slightly outperformed shared budgets at this sparsity and had lower QK error at all three tabulated sparsities; shared budgets achieved lower PPL at 30% and 50%. Thus, freeing the per-row quota mattered substantially here, while sharing the budget between Q and K was not uniformly better than separate matrix budgets.

Effect of RoPE. Adding RoPE did not consistently improve perplexity. At matched budgets, it lowered WikiText-2 PPL at 3 of 9 nonzero sparsities for separate budgets and 2 of 9 for shared budgets; on C4, it improved 3 of 9 for each budget. Below 40% sparsity, all absolute changes were smaller than 0.001 PPL. From 40% through 80%, RoPE increased PPL on both datasets for both budgets. At 90%, however, it reduced WikiText-2/C4 PPL from 29.23/40.29 to 26.66/36.40 with separate budgets and from 30.14/41.90 to 27.19/37.14 with a shared budget. Thus, RoPE helped at the most aggressive pruning level, despite being slightly worse over the intermediate range.

Zero-shot changes were small and mixed. At a matched shared budget, adding RoPE changed mean accuracy by -0.03, $+0.10$, and $+0.01$ percentage points at 30%, 50%, and 80% sparsity, respectively. At 80%, RoPE with separate budgets achieved 58.13%, compared with 57.82% for RoPE with a shared budget. Both RoPE variants had higher pre-RoPE QK error than the shared causal method at all three tabulated sparsities, while retaining lower error than Wanda. This comparison concerns a common pre-RoPE diagnostic, not whether the RoPE-aware objective was better preserved. These single-seed results support retaining the simpler causal pre-RoPE score as the default for this setup, while identifying a benefit from RoPE at 90% sparsity.

H.3 TINYLLAMA 1.1B

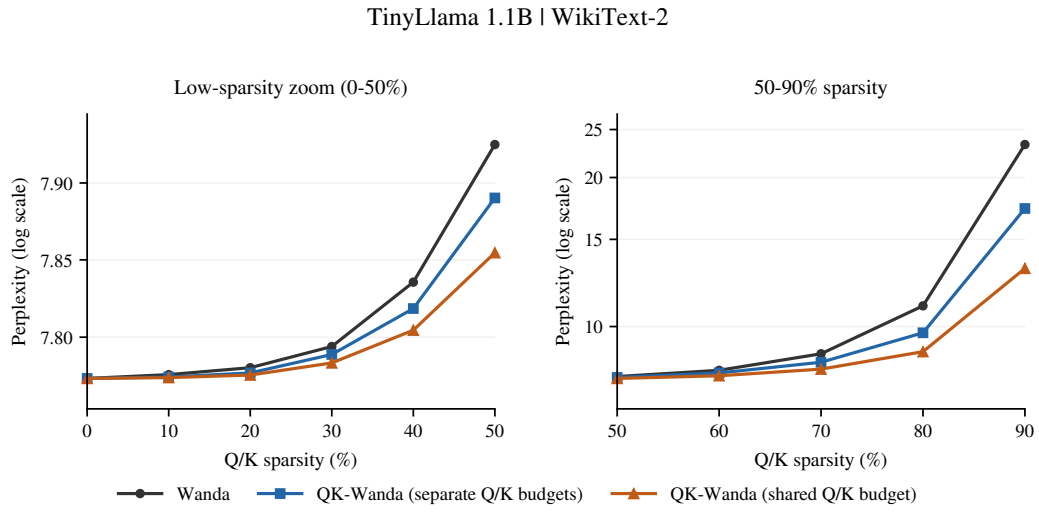

Figure 10: TinyLlama-1.1B on WikiText-2. The 0–50% and 50–90% panels compare the same three methods using independent logarithmic PPL ranges.

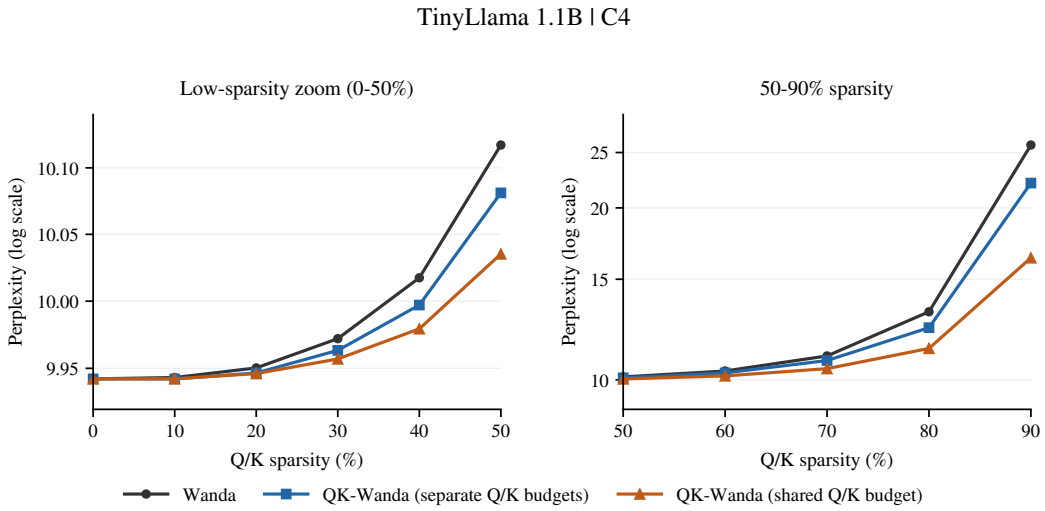

Figure 11: TinyLlama-1.1B on C4, under the same pruning and calibration protocol as Figure 10.

H.4 LLAMA 3.2 1B

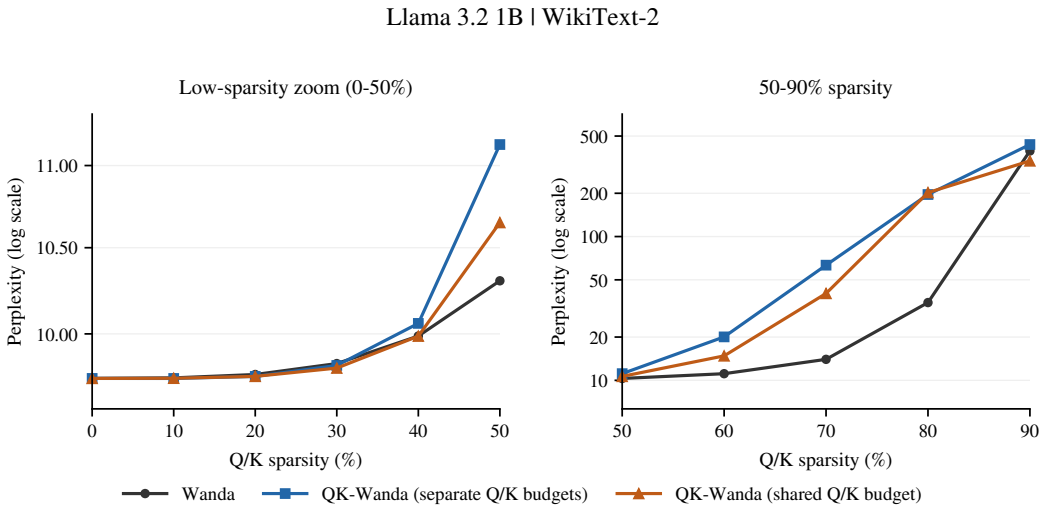

Figure 12: Llama-3.2-1B on WikiText-2. The 0–50% and 50–90% panels compare the same three methods using independent logarithmic PPL ranges.

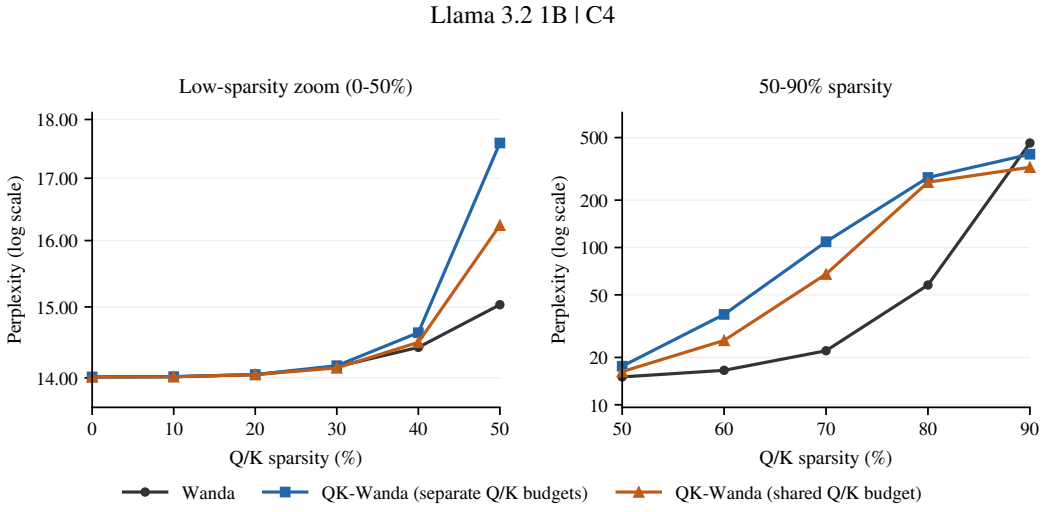

Figure 13: Llama-3.2-1B on C4, under the same pruning and calibration protocol as Figure 12.

H.5 LLAMA 3.2 3B

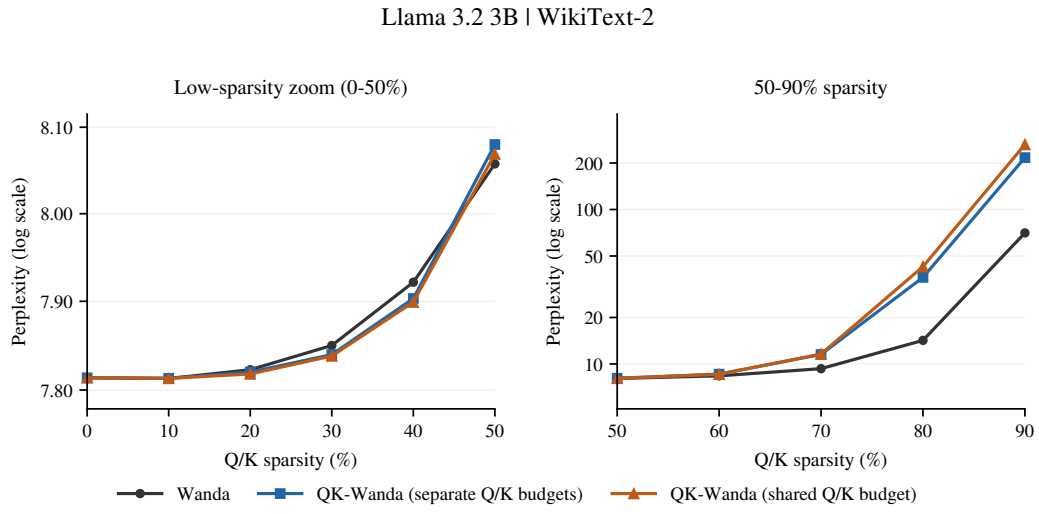

Figure 14: Llama-3.2-3B on WikiText-2. The 0–50% and 50–90% panels compare the same three methods using independent logarithmic PPL ranges.

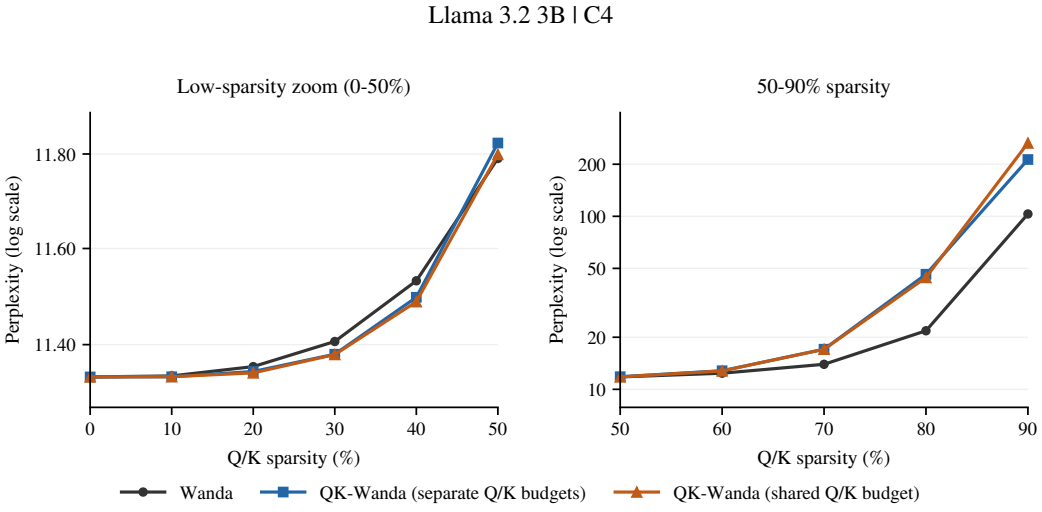

Figure 15: Llama-3.2-3B on C4, under the same pruning and calibration protocol as Figure 14.

H.6 LLAMA 3.1 8B

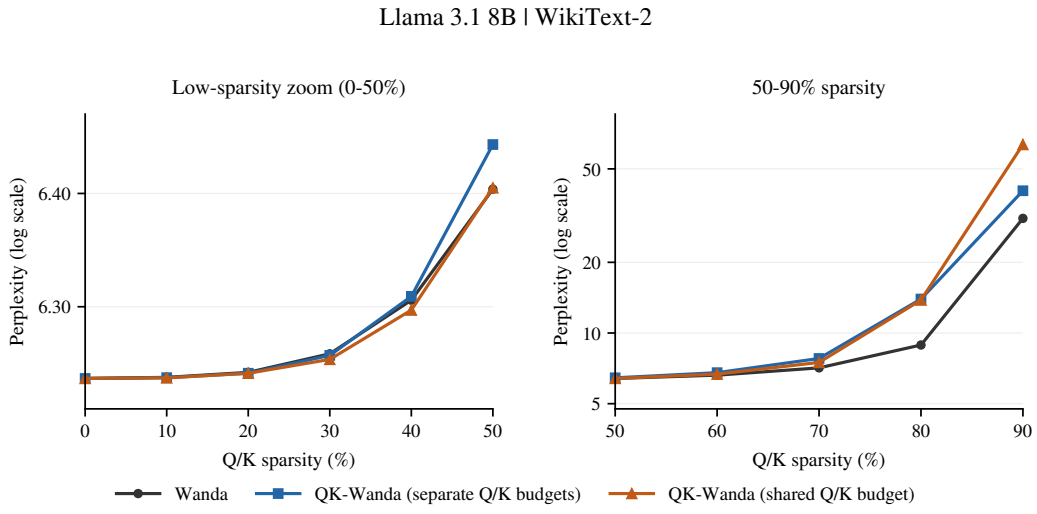

Figure 16: Llama-3.1-8B on WikiText-2. The 0–50% and 50–90% panels compare the same three methods using independent logarithmic PPL ranges.

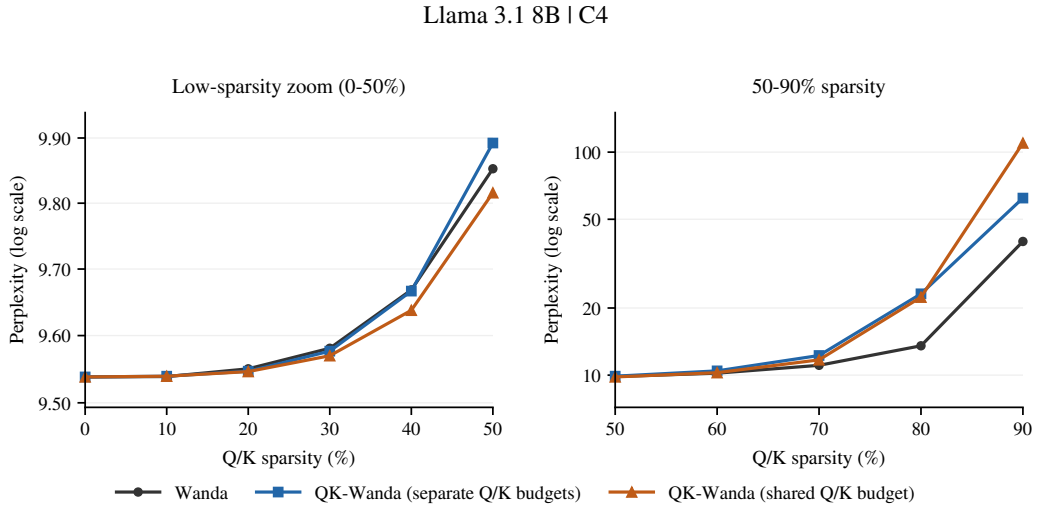

Figure 17: Llama-3.1-8B on C4, under the same pruning and calibration protocol as Figure 16.

H.7 QWEN2.5

We used the same two-panel presentation for the Qwen2.5 comparisons in Section F.3: 0–50% sparsity on the left and 50–90% on the right, with independent logarithmic PPL ranges and measurements ending at 90%. Each figure compares original row-wise WANDA, separate-matrix QK-WANDA, and shared-budget QK-WANDA.

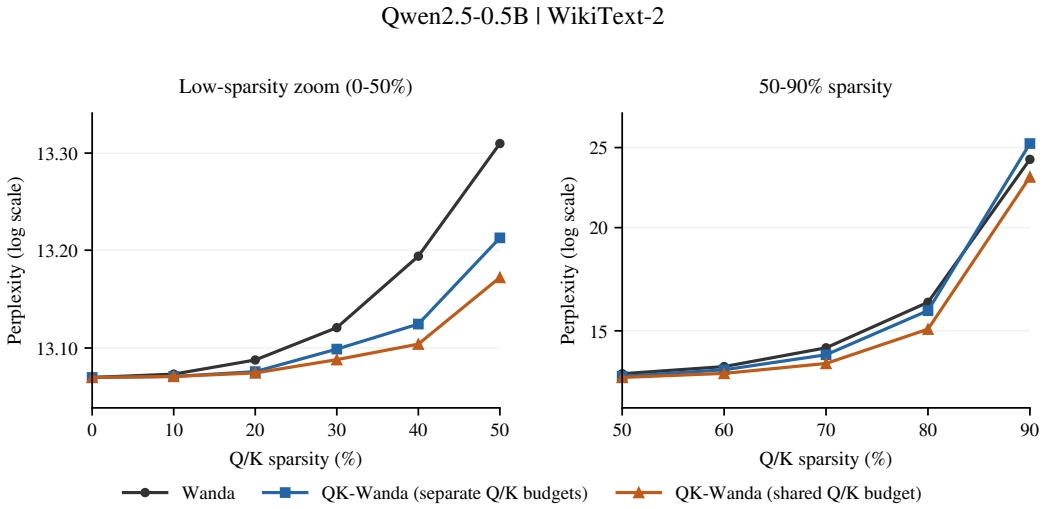

Figure 18: Qwen2.5-0.5B on WikiText-2. The 0–50% and 50–90% panels compare the three budget methods under the protocol in Section F.3, using independent logarithmic PPL ranges.

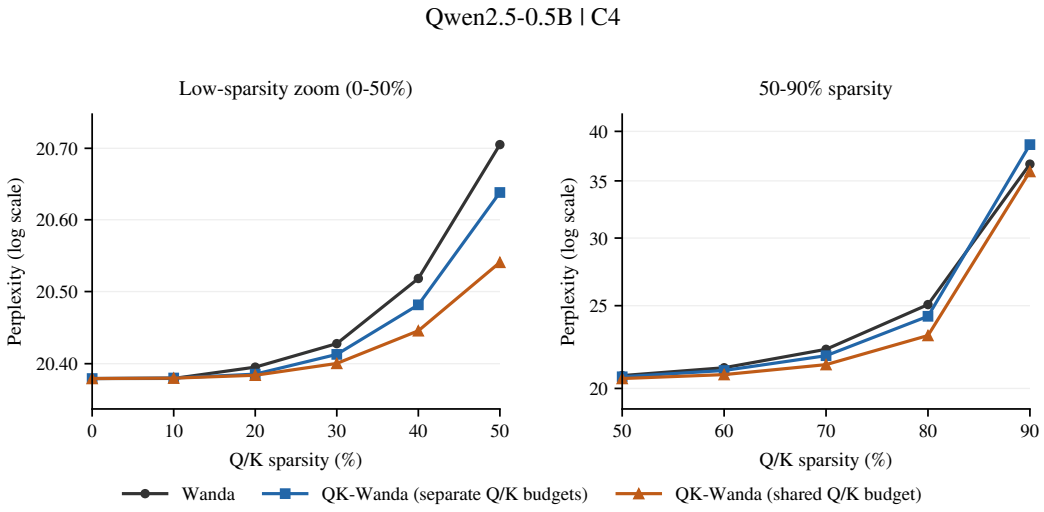

Figure 19: Qwen2.5-0.5B on C4. The 0–50% and 50–90% panels compare the three budget methods under the protocol in Section F.3, using independent logarithmic PPL ranges.

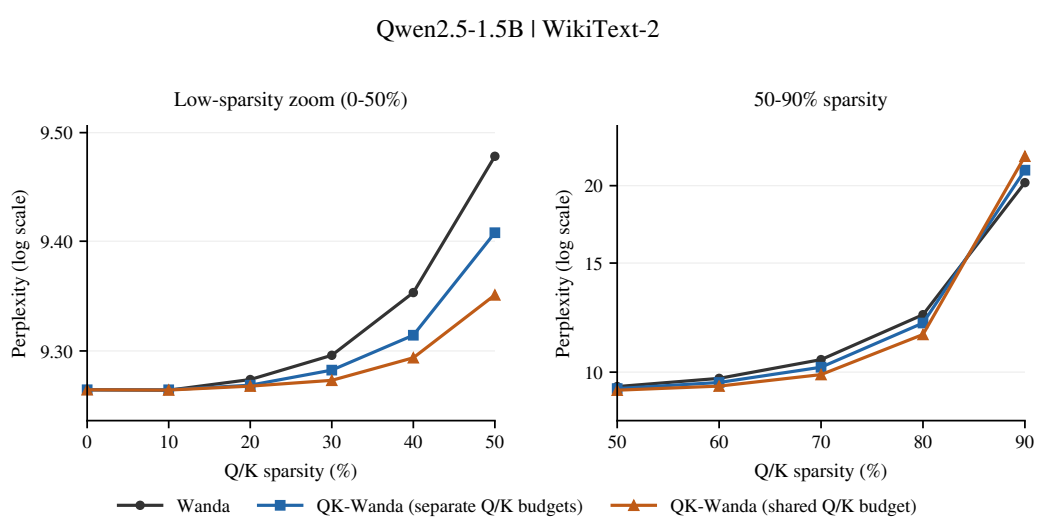

Figure 20: Qwen2.5-1.5B on WikiText-2. The 0–50% and 50–90% panels compare the three budget methods under the protocol in Section F.3, using independent logarithmic PPL ranges.

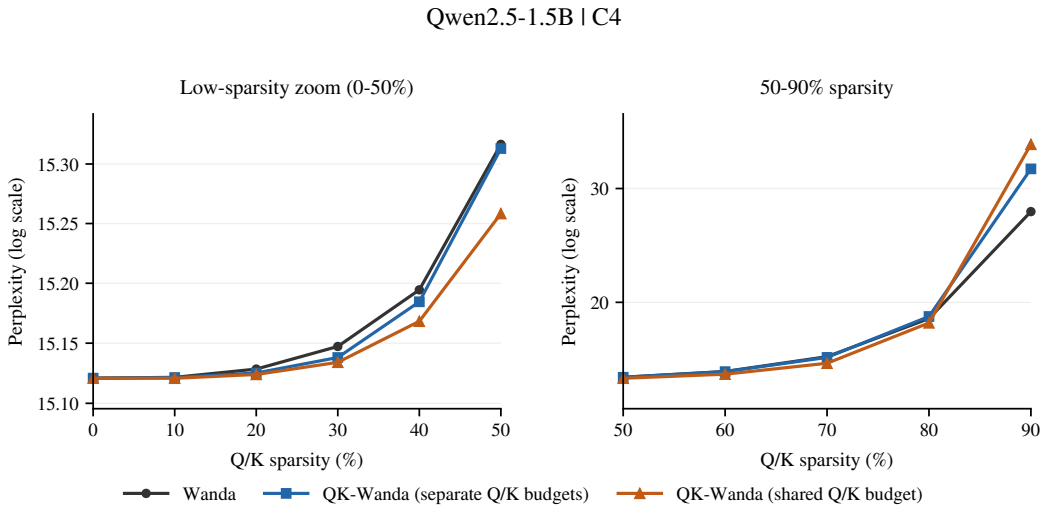

Figure 21: Qwen2.5-1.5B on C4. The 0–50% and 50–90% panels compare the three budget methods under the protocol in Section F.3, using independent logarithmic PPL ranges.

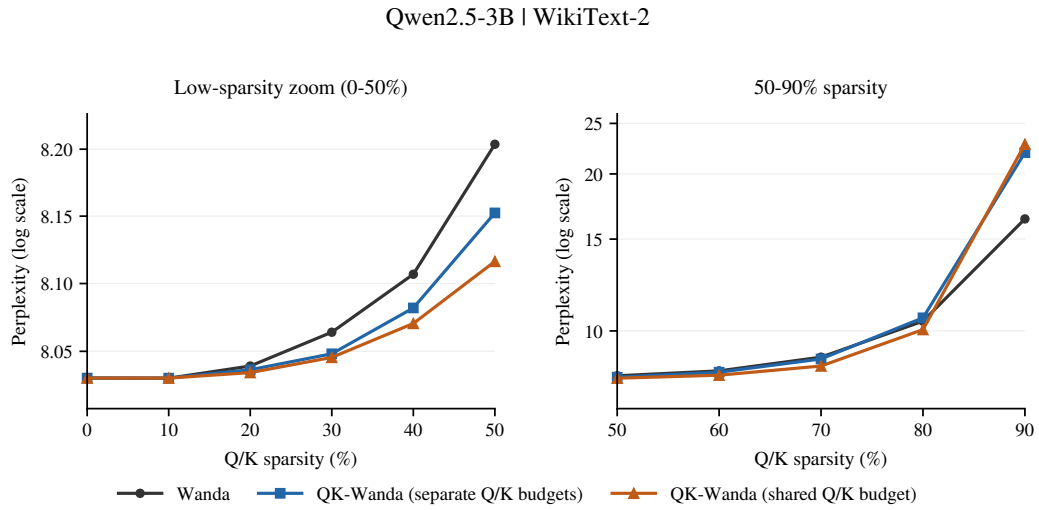

Figure 22: Qwen2.5-3B on WikiText-2. The 0–50% and 50–90% panels compare the three budget methods under the protocol in Section F.3, using independent logarithmic PPL ranges.

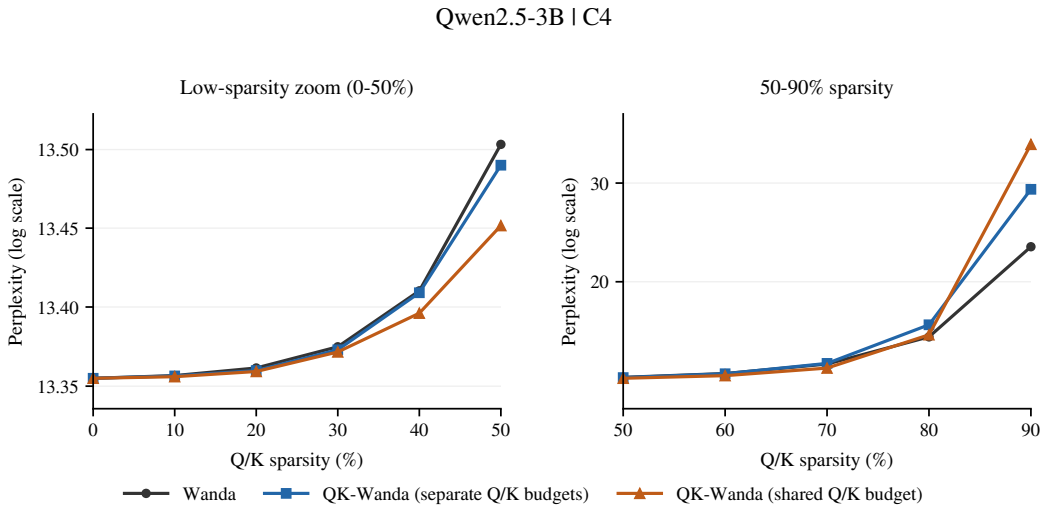

Figure 23: Qwen2.5-3B on C4. The 0–50% and 50–90% panels compare the three budget methods under the protocol in Section F.3, using independent logarithmic PPL ranges.

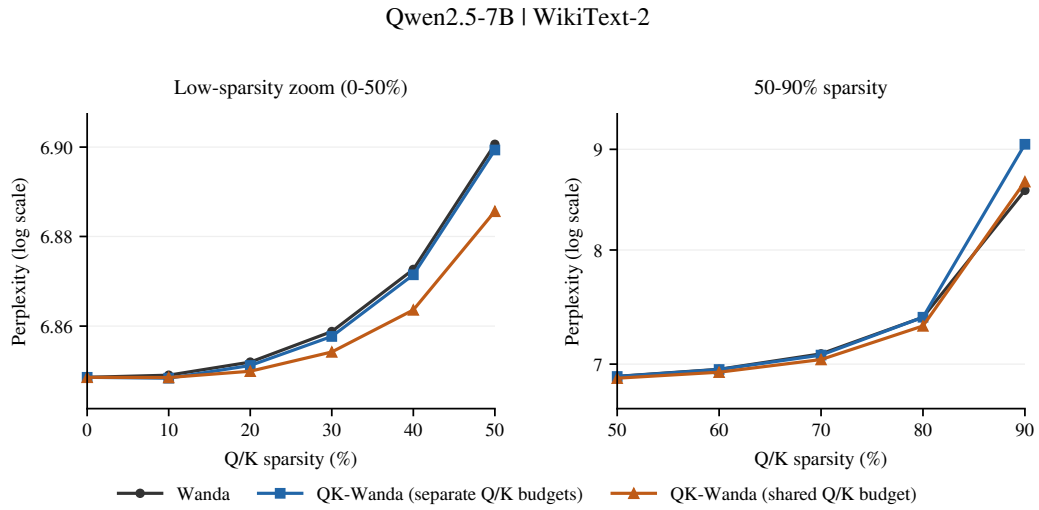

Figure 24: Qwen2.5-7B on WikiText-2. The 0–50% and 50–90% panels compare the three budget methods under the protocol in Section F.3, using independent logarithmic PPL ranges.

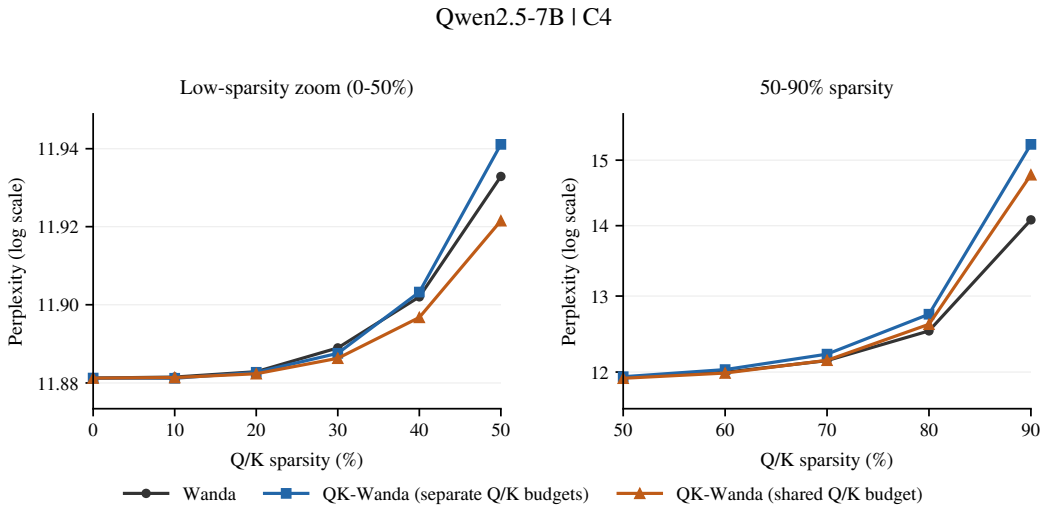

Figure 25: Qwen2.5-7B on C4. The 0–50% and 50–90% panels compare the three budget methods under the protocol in Section F.3, using independent logarithmic PPL ranges.

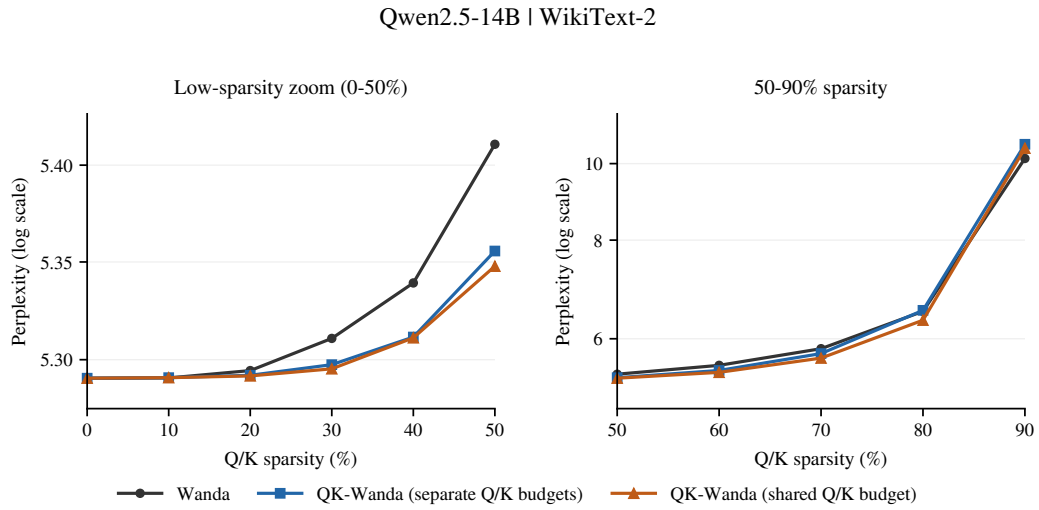

Figure 26: Qwen2.5-14B on WikiText-2. The 0–50% and 50–90% panels compare the three budget methods under the protocol in Section F.3, using independent logarithmic PPL ranges.

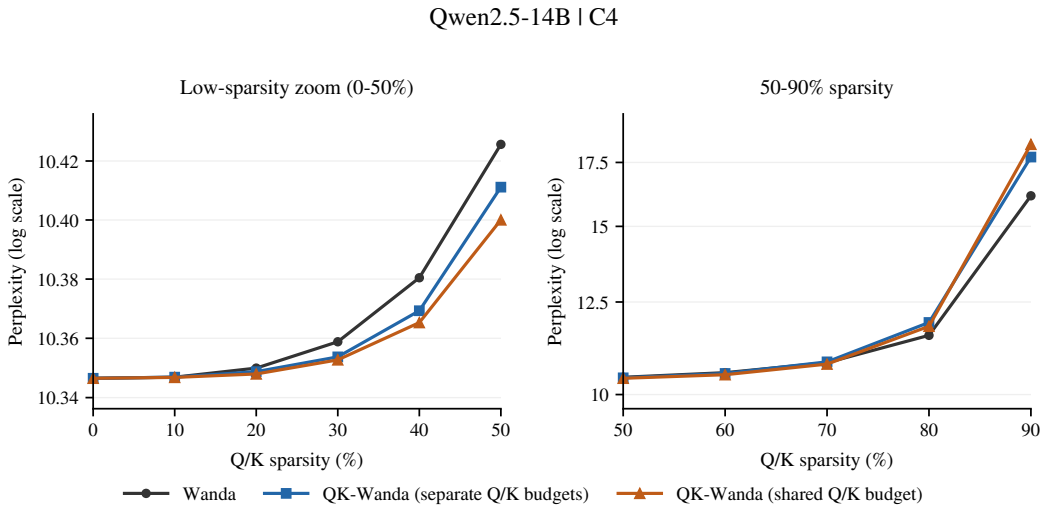

Figure 27: Qwen2.5-14B on C4. The 0–50% and 50–90% panels compare the three budget methods under the protocol in Section F.3, using independent logarithmic PPL ranges.

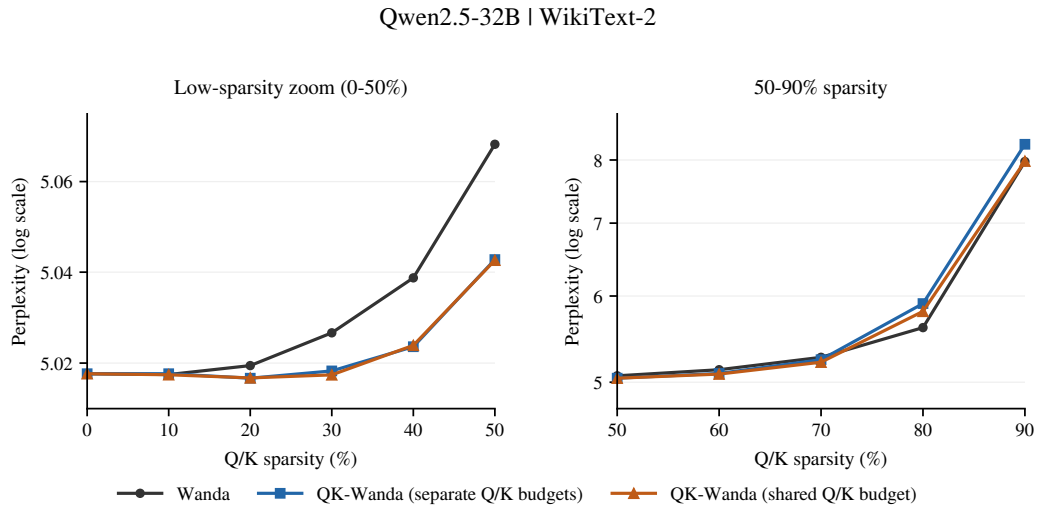

Figure 28: Qwen2.5-32B on WikiText-2. The 0–50% and 50–90% panels compare the three budget methods under the protocol in Section F.3, using independent logarithmic PPL ranges.

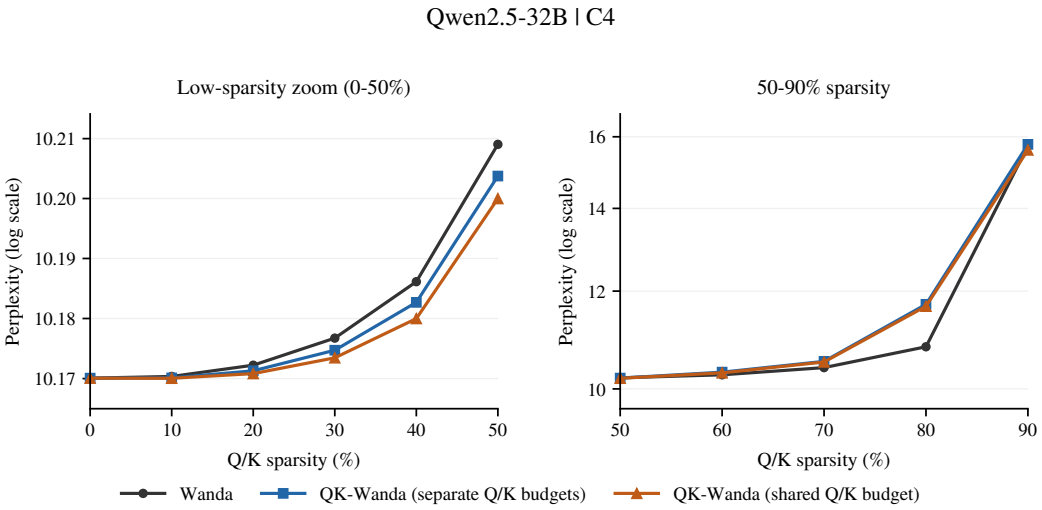

Figure 29: Qwen2.5-32B on C4. The 0–50% and 50–90% panels compare the three budget methods under the protocol in Section F.3, using independent logarithmic PPL ranges.

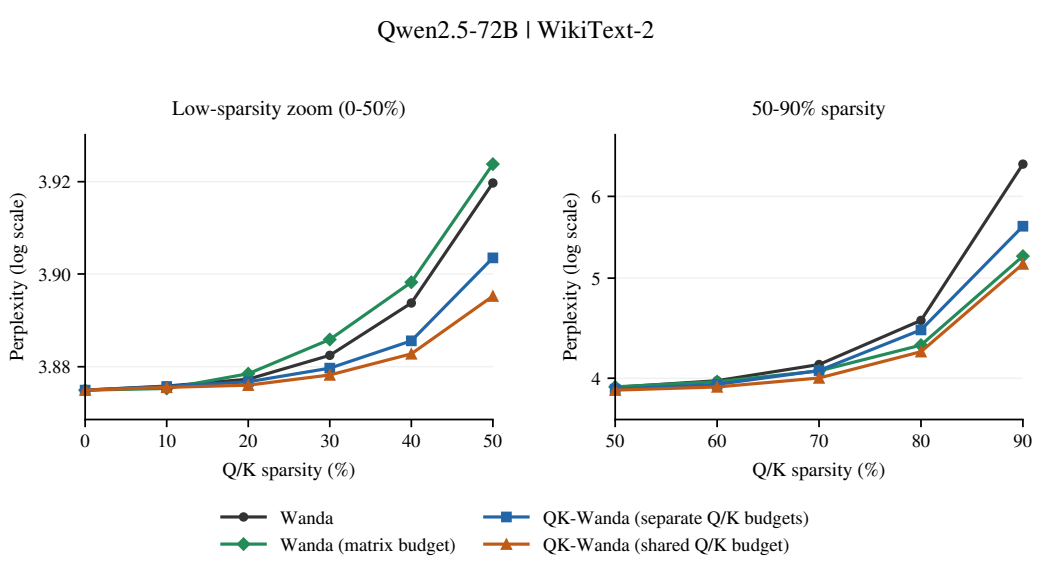

Figure 30: Qwen2.5-72B on WikiText-2. The 0–50% and 50–90% panels compare original row-wise WANDA, its matrix-budget control, and separate/shared QK-WANDA, with independent logarithmic PPL ranges. All four methods and the dense reference use BF16 weights under the protocol in Section F.3.

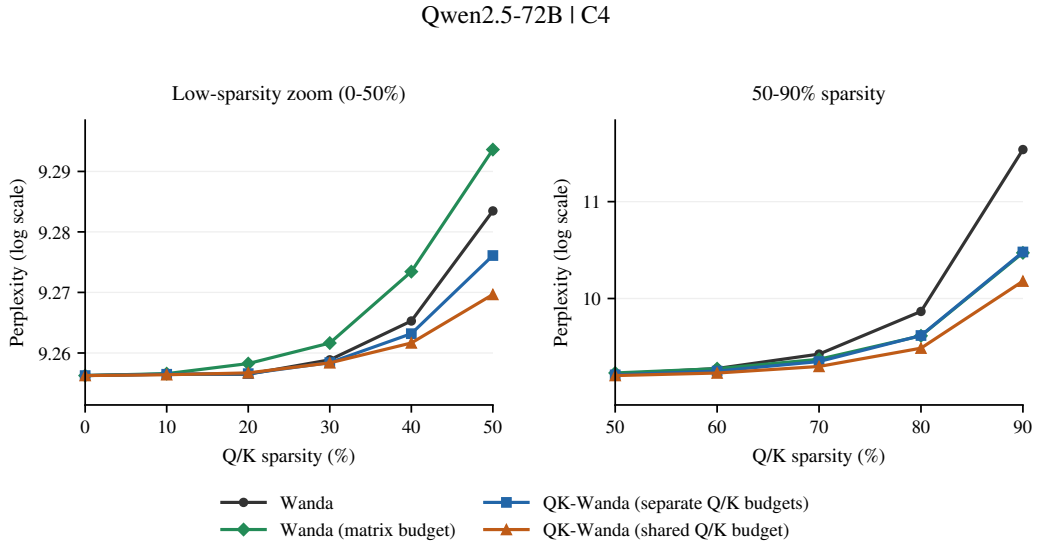

Figure 31: Qwen2.5-72B on C4. The 0–50% and 50–90% panels use the same four methods, BF16 checkpoint, and calibration as Figure 30, with independent logarithmic PPL ranges.

H.8 LLAMA 2 70B

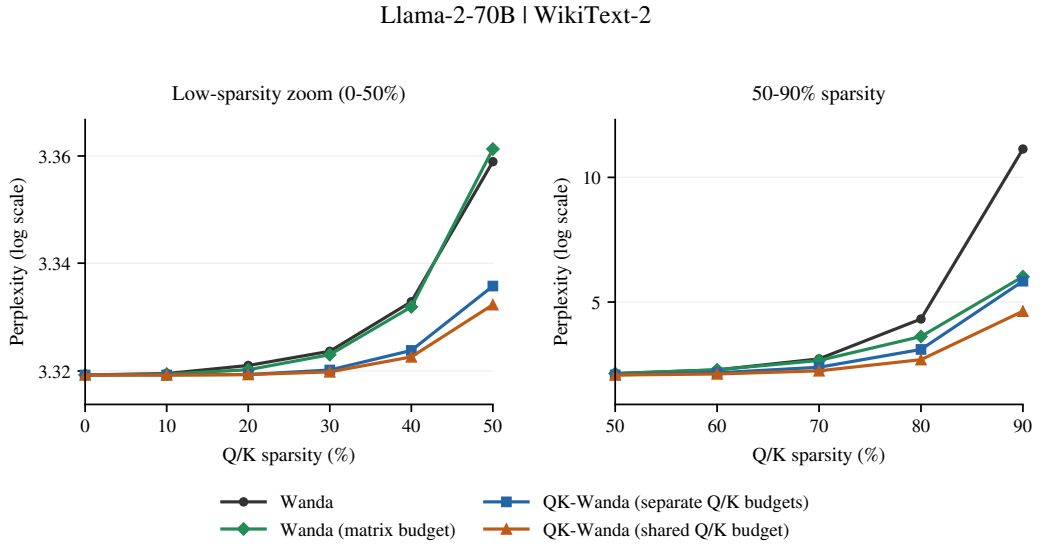

Figure 32: Llama 2 70B on WikiText-2. Only Q/K weights were pruned, using the common calibration and evaluation protocol in Section 4. Original row-wise WANDA, its matrix-budget control, and separate/shared QK-Wanda budgets are compared over 0–50% and 50–90% sparsity, with independent logarithmic PPL ranges.

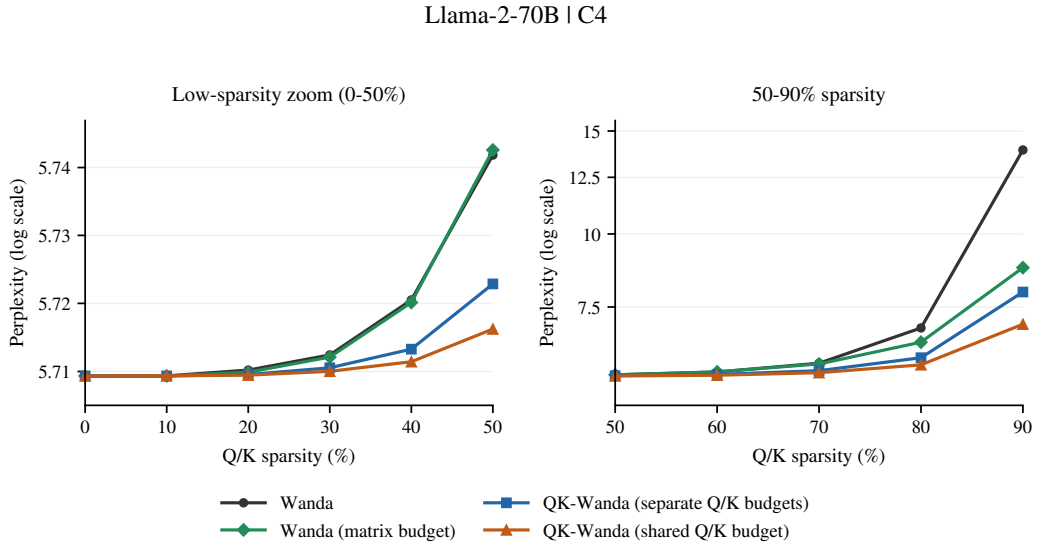

Figure 33: Llama 2 70B on C4. Only Q/K weights were pruned, using the common calibration and evaluation protocol in Section 4. Original row-wise WANDA, its matrix-budget control, and separate/shared QK-Wanda budgets are compared over 0–50% and 50–90% sparsity, with independent logarithmic PPL ranges.

H.9 LLAMA 3.1 70B

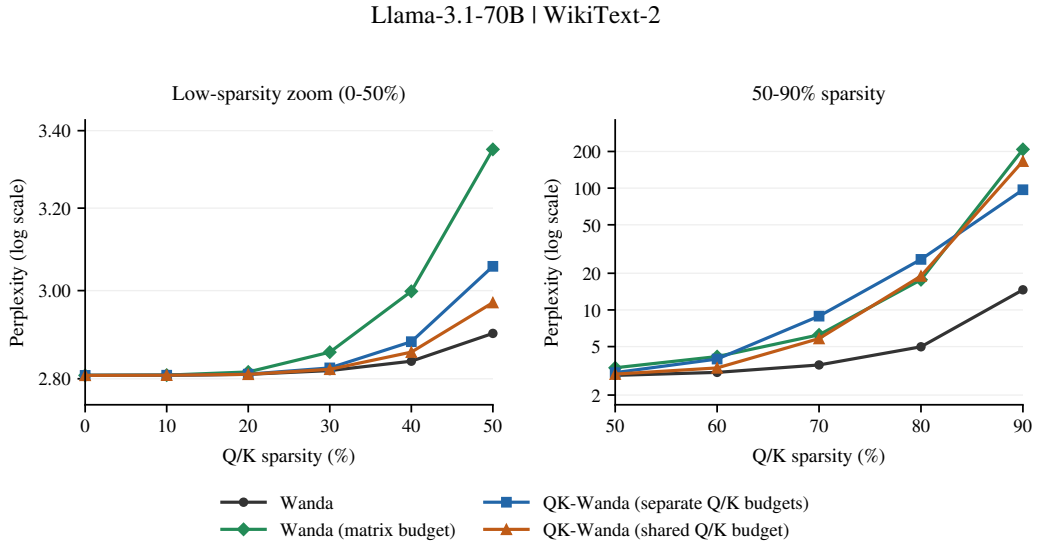

Figure 34: Llama 3.1 70B on WikiText-2. Only Q/K weights were pruned, using the common calibration and evaluation protocol in Section 4. Original row-wise WANDA, its matrix-budget control, and separate/shared QK-Wanda budgets are compared over 0–50% and 50–90% sparsity, with independent logarithmic PPL ranges.

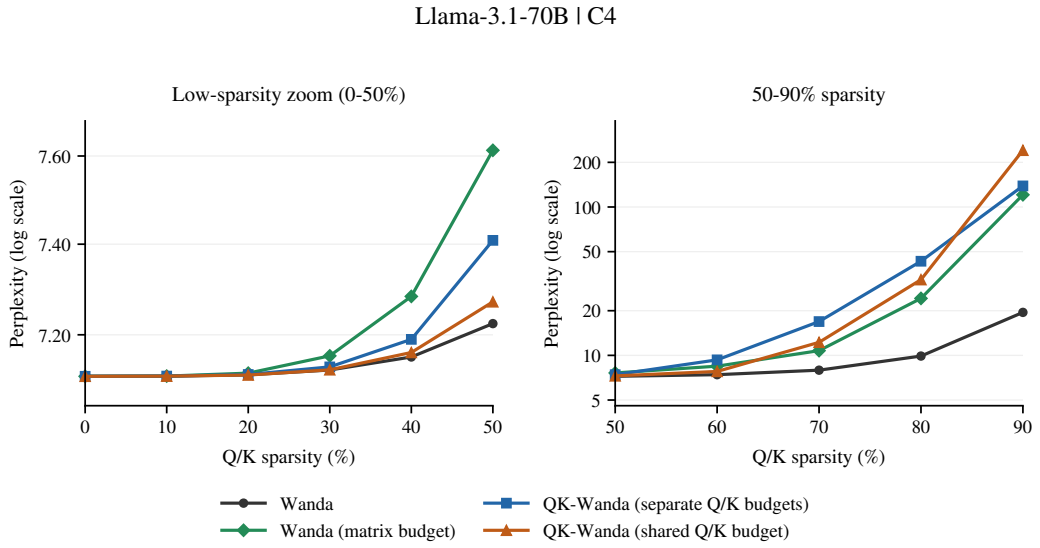

Figure 35: Llama 3.1 70B on C4. Only Q/K weights were pruned, using the common calibration and evaluation protocol in Section 4. Original row-wise WANDA, its matrix-budget control, and separate/shared QK-Wanda budgets are compared over 0–50% and 50–90% sparsity, with independent logarithmic PPL ranges.

I QK RECONSTRUCTION ACROSS SPARSITY LEVELS

We evaluated the relative QK error in Equation (47) for all 15 checkpoints at 10%, 20%, . . . , 90% Q/K sparsity. Each comparison used the exact masks from the corresponding PPL sweep: original row-wise WANDA, separate-matrix QK-WANDA, and shared-budget QK-WANDA. We retained the 16 held-out C4 sequences of length 512 and the common dense block inputs described in Appendix D.3. The previously reported Wanda/shared values at 30%, 50%, and 80% were reused; the remaining 315 configurations were measured under the same per-checkpoint protocol. A repeated 50% shared-budget measurement for every checkpoint reproduced the original per-block errors.

Each panel uses a logarithmic vertical axis for the entire measured sparsity range. Positive errors spanned several orders of magnitude, making this scale more informative than a linear axis at low sparsity. Dense reconstruction error is exactly zero and is omitted from the logarithmic axis; it was not replaced with a positive plotting constant. Markers indicate measured configurations, and lines connect adjacent measurements. Vertical ranges vary across checkpoints.

Measured reconstruction differences. Shared-budget QK-WANDA had lower QK error than WANDA in 133 of 135 checkpoint–sparsity comparisons; the separate-budget variant did so in 124 of 135. Shared budgets gave lower error than separate budgets in 121 of 135 comparisons. These are descriptive counts for the fixed calibration seed and held-out set, not independent statistical trials.

The two QK allocations produced similar curves on Llama 2 7B and 13B. On each of these checkpoints, separate budgets had slightly lower error at seven of the nine nonzero sparsities. On Llama-3.2-3B, shared budgets produced higher error than Wanda at high sparsity: at 80%/90%, shared-budget errors were 9.80%/24.10%, versus 9.58%/20.68% for Wanda. The relation between this local diagnostic and downstream PPL is discussed in Appendix D.3.

Largest checkpoints. At 90% sparsity, moving from Wanda to shared-budget QK-Wanda reduced relative QK error on Llama 2 70B (17.47% to 3.01%), Llama 3.1 70B (11.12% to 7.35%), and Qwen2.5-72B (2.37% to 1.28%).

I.1 LLAMA 2 AND TINYLLAMA

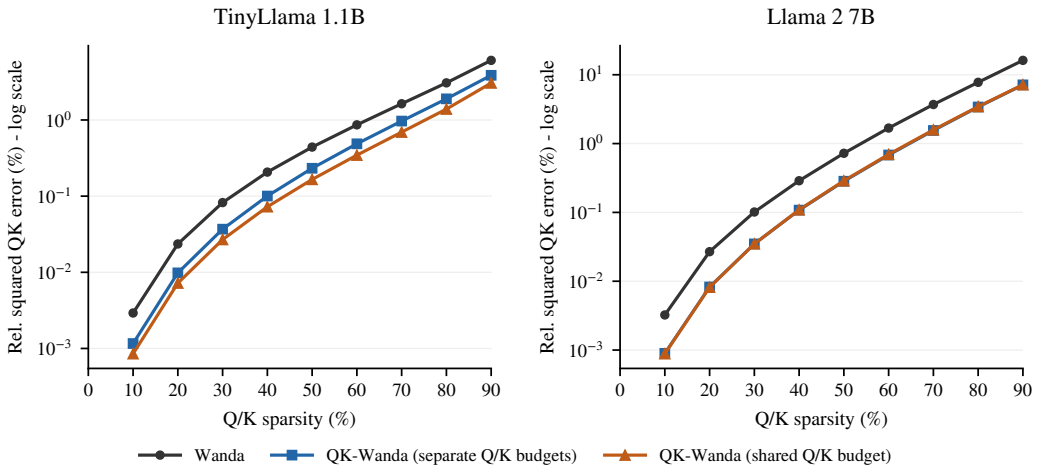

Figure 36: TinyLlama 1.1B (left) and Llama 2 7B (right): relative squared causal pre-RoPE QK error on held-out C4 inputs, using the exact masks from PPL evaluation. Dense error is zero and is omitted from the logarithmic axes.

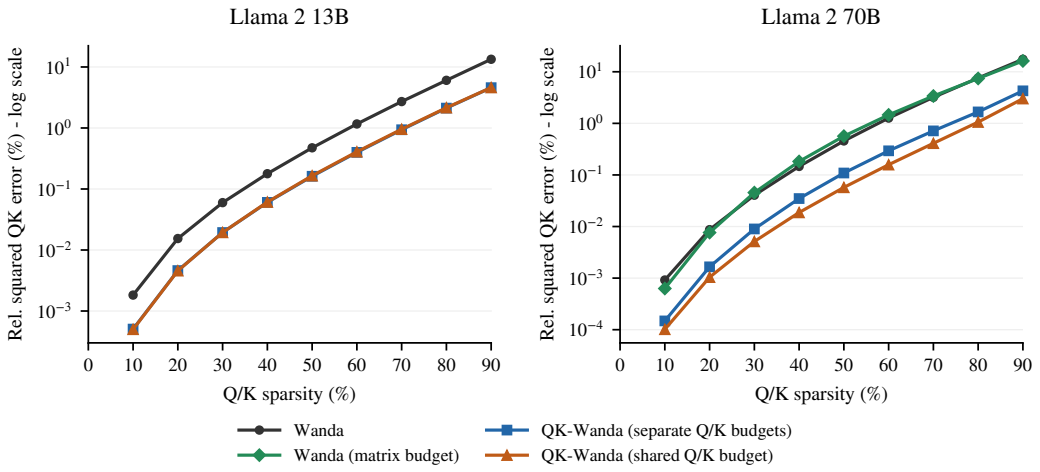

Figure 37: Llama 2 13B (left) and Llama 2 70B (right): relative squared causal pre-RoPE QK error on held-out C4 inputs, using the exact masks from PPL evaluation. Dense error is zero and is omitted from the logarithmic axes. The matrix-budget Wanda control is included for Llama 2 70B.

I.2 LLAMA 3

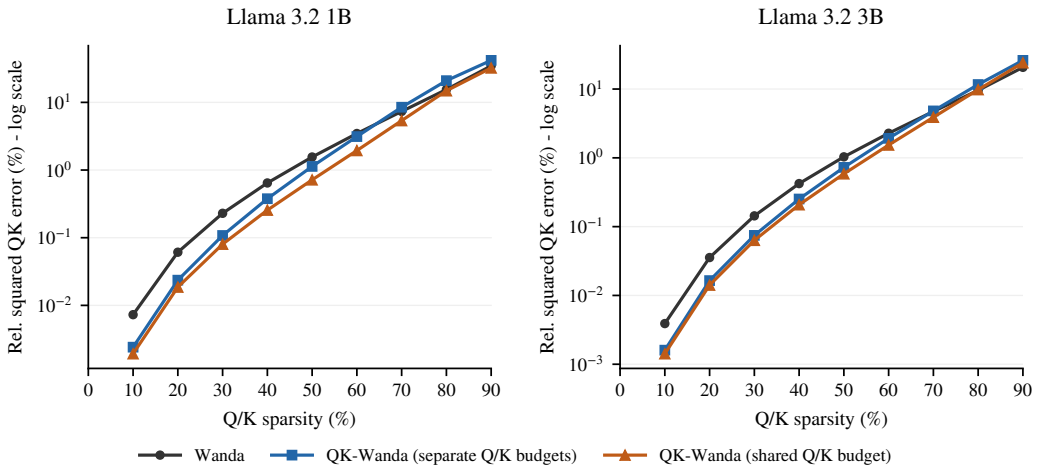

Figure 38: Llama 3.2 1B (left) and Llama 3.2 3B (right): relative squared causal pre-RoPE QK error on held-out C4 inputs, using the exact masks from PPL evaluation. Dense error is zero and is omitted from the logarithmic axes.

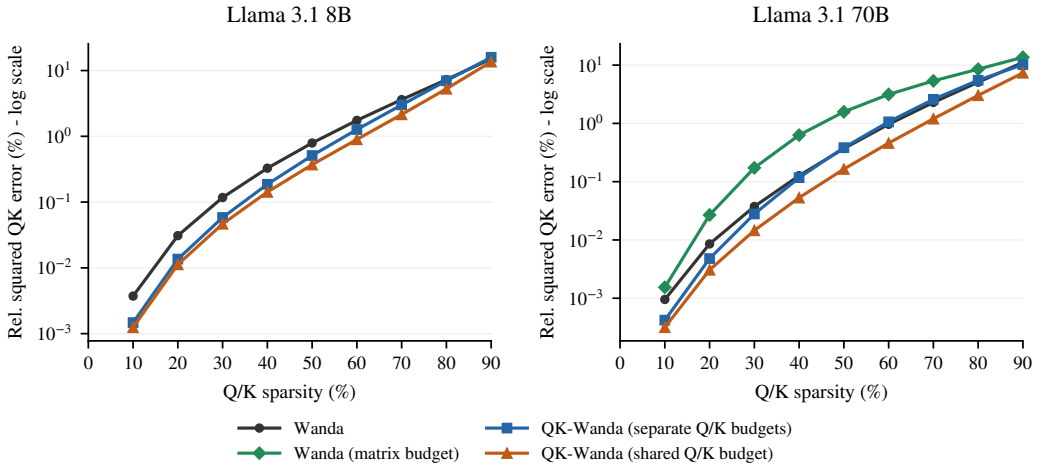

Figure 39: Llama 3.1 8B (left) and Llama 3.1 70B (right): relative squared causal pre-RoPE QK error on held-out C4 inputs, using the exact masks from PPL evaluation. Dense error is zero and is omitted from the logarithmic axes. The matrix-budget Wanda control is included for Llama 3.1 70B.

I.3 QWEN2.5

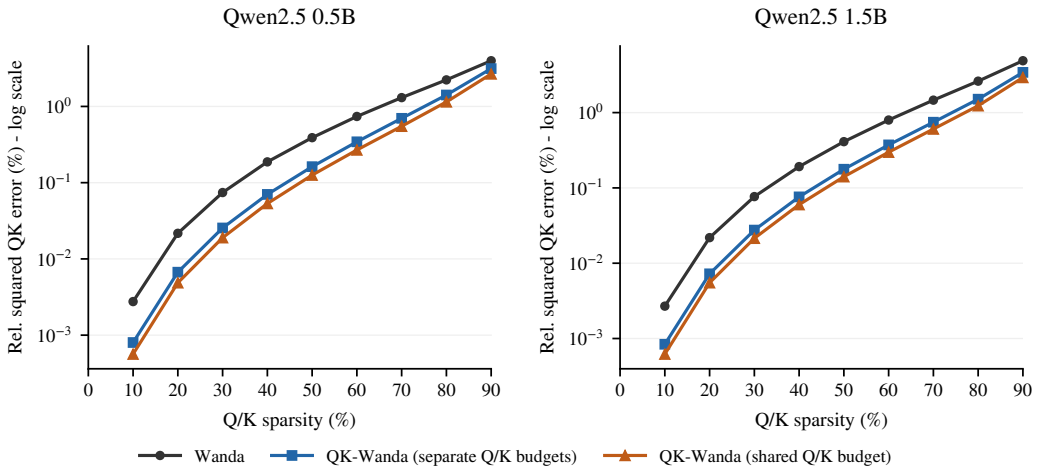

Figure 40: Qwen2.5 0.5B (left) and Qwen2.5 1.5B (right): relative squared causal pre-RoPE QK error on held-out C4 inputs, using the exact masks from PPL evaluation. Dense error is zero and is omitted from the logarithmic axes.

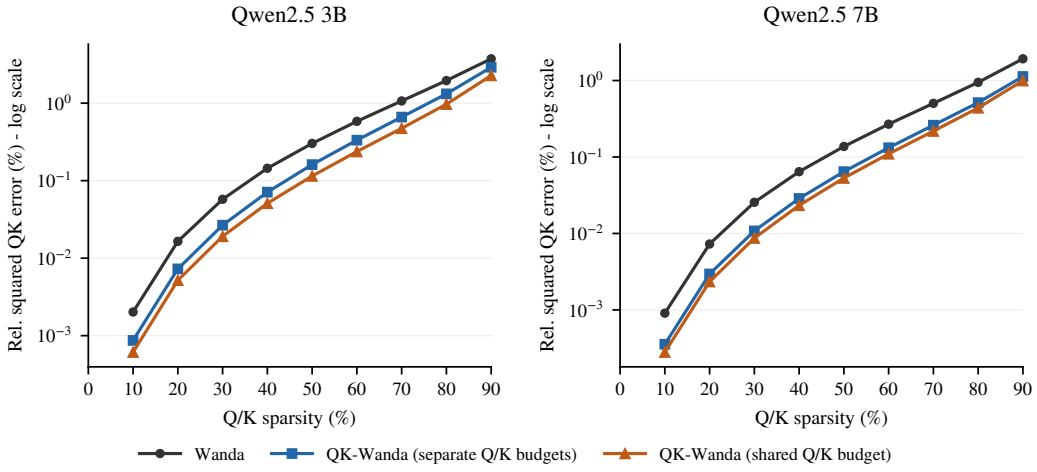

Figure 41: Qwen2.5 3B (left) and Qwen2.5 7B (right): relative squared causal pre-RoPE QK error on held-out C4 inputs, using the exact masks from PPL evaluation. Dense error is zero and is omitted from the logarithmic axes.

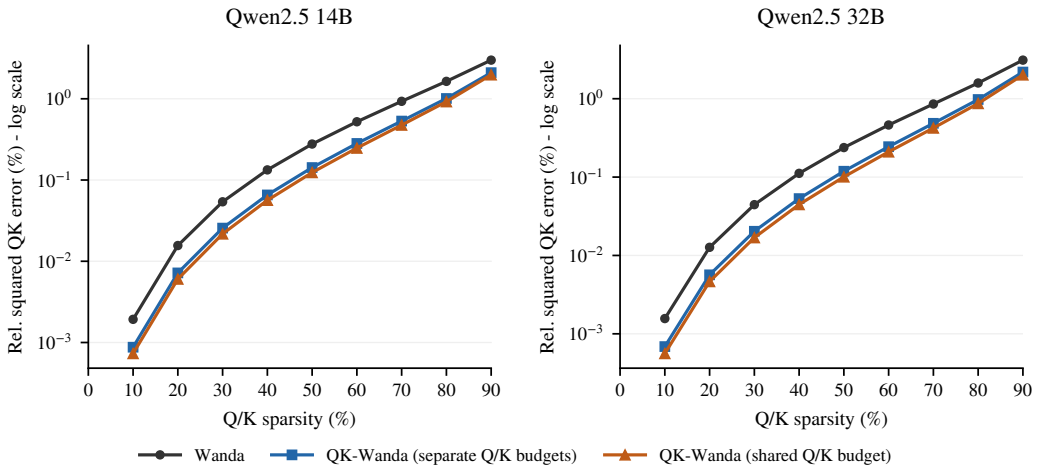

Figure 42: Qwen2.5 14B (left) and Qwen2.5 32B (right): relative squared causal pre-RoPE QK error on held-out C4 inputs, using the exact masks from PPL evaluation. Dense error is zero and is omitted from the logarithmic axes.

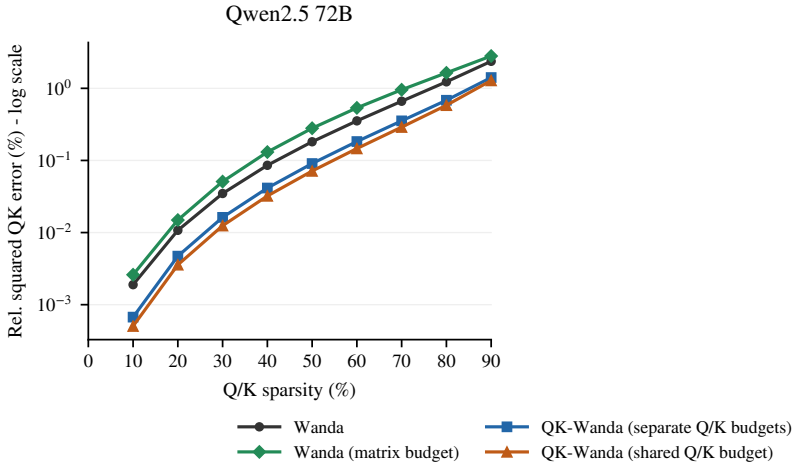

Figure 43: Qwen2.5 72B: relative squared causal pre-RoPE QK error on held-out C4 inputs, using the exact masks from PPL evaluation. Dense error is zero and is omitted from the logarithmic axes. The matrix-budget Wanda control is included for Qwen2.5 72B.

J PROGRESSIVE Q/K RE-SCORING

J.1 MOTIVATION AND THE EXECUTED SCHEDULE

The QK-WANDA score of a query weight depends on keys, and a key score depends on queries. Partial pruning therefore changes the energies used by subsequent decisions. We tested whether refreshing those energies while gradually growing the masks improves the final model.

Let J be the number of steps and $M = \lfloor 0.5n \rfloor$ the final removal count in a group of n original weights. A group is one output row in the row-wise setting, or an entire projection matrix in the matrix-global setting. At step $r = 1, \ldots, J$, both QK-WANDA score arrays are computed from the current sparse pair $(W^{Q,r-1}, W^{K,r-1})$, where the second superscript entry counts completed pruning steps. With initial removal count $M_0 = 0$, the cumulative target is

$$M_r = \left\lfloor \frac{rM}{J} \right\rfloor . \tag{50}$$

We add the $M_r - M_{r-1}$ lowest-scoring unmasked coordinates, separately for Q and K. Previously deleted entries stay deleted. Both next masks are constructed before either is applied. Inputs to the current block remain fixed during these steps; its final sparse output feeds the next block.

This is simultaneous progressive mask growth, rather than an alternating Q-then-K update. For $J = 5$, cumulative sparsity is approximately 10%, 20%, 30%, 40%, 50%; $J = 10$ uses approximately 5% increments. $J = 1$ is one-shot QK-WANDA. No retained weight is updated by compensation.

J.2 WHY RE-SCORING IS NOT A GREEDY TEACHER-LOSS OPTIMIZER

The scalar proof applies around the *current* pair. It measures the squared local change caused by the next deletion, rather than its marginal effect on reconstruction of the original reference. Using the unscaled products $B = Q^\top K$ from Section A.1, for each sequence and head let $E = B_{\text{current}} - B_{\text{ref}}$ be the accumulated error, and let $D = B_{\text{next}} - B_{\text{current}}$ be the next change. The new error is $E + D$; we suppress sequence/head labels in these three expressions. For two collections of head matrices $A = \{A_h^{(s)}\}$ and $B = \{B_h^{(s)}\}$, define the masked sum of products

$$\langle A, B \rangle_C = \frac{1}{N} \sum_{s=1}^{N} \sum_{h=1}^{H_Q} \sum_{t=1}^{T_s} \sum_{u=1}^{t} A_{h,tu}^{(s)} B_{h,tu}^{(s)}. \tag{51}$$

The brackets $\langle \cdot, \cdot \rangle_C$ are shorthand for this expression, and $\|A\|_C^2 = \langle A, A \rangle_C$ is its associated squared masked error. The subscript C records the causal restriction. Expanding the square gives the exact change in the causal loss $\mathcal{L}_C$ against the fixed reference (the "teacher"):

$$\|E + D\|_C^2 - \|E\|_C^2 = 2\langle E, D \rangle_C + \|D\|_C^2 . \tag{52}$$

The refreshed scalar score accounts for $\|D\|_C^2$ for one deletion; it omits the residual inner product. Simultaneous batches also contain the interactions in Equation (28). Thus these experiments test refreshed scalar ranking, not exact greedy minimization, residual-aware selection, or compensated pruning.

J.3 RESULTS AND SOURCE PROVENANCE

The ablation used Llama-3.2-1B, Q/K-only unstructured pruning to 50%, and calibration seed 0. C4 calibration uses 128×2048 or 256×32 tokens; PPL uses length 2048. Times cover pruning on one A100 40GB, excluding evaluation.

Table 19: Row-wise progressive pruning on Llama-3.2-1B. The final budget is 50% in every output row. Each row is one seed-0 run; bold selects the lowest unrounded PPL for each calibration setting.

Calib. $N \times T$	Score	Steps	WikiText-2 PPL	C4 PPL	Time (s)
128×2048	WANDA	–	10.6474	15.7481	22.2
128×2048	QK-WANDA	1	10.5334	**15.5175**	26.2
128×2048	QK-WANDA	10	**10.5332**	15.5178	94.9
256×32	WANDA	–	**10.3054**	15.0255	15.7
256×32	QK-WANDA	1	10.3072	**15.0187**	17.2
256×32	QK-WANDA	2	10.3077	15.0193	21.9
256×32	QK-WANDA	5	10.3066	15.0193	39.7
256×32	QK-WANDA	10	10.3067	15.0195	65.7

Table 20: Progressive QK-Wanda pruning with separate matrix budgets. Each Q and K matrix has its own 50% budget. All other settings match Table 19; this ablation used a separate single-seed calibration protocol. Bold marks the lowest unrounded PPL for each calibration setting.

Calib. $N \times T$	Score	Steps	WikiText-2 PPL	C4 PPL	Time (s)
128×2048	QK-WANDA	1	**11.4888**	**18.2040**	25.9
128×2048	QK-WANDA	10	12.5458	20.8593	94.9
256×32	QK-WANDA	1	**11.1735**	**17.7755**	19.8
256×32	QK-WANDA	2	11.2330	17.9338	22.4
256×32	QK-WANDA	5	11.4152	18.5233	37.0
256×32	QK-WANDA	10	11.6609	19.1525	66.3

The one-shot rows also compare scoring rules at fixed allocation granularity. With 128×2048 calibration, causal QK-Wanda improves both PPL values over Wanda; with 256×32, it slightly worsens WikiText-2 and improves C4. Causal weighting can change rankings within a row (Appendix A.3), unlike the single-sequence unmasked row rescaling. These controls show a calibration-dependent scoring effect at one checkpoint and sparsity; they do not isolate how much of the main cross-model gains comes from scoring versus allocation.

Under row budgets, QK-WANDA schedules differ by less than 0.0012 PPL within each calibration setting, while ten-step pruning takes 3.6–3.8 times as long as one-shot QK-WANDA. These single-seed results provide no evidence of a useful quality gain.

Under matrix-global budgets, additional steps worsen both perplexities. At length 32, one to ten steps raises WikiText-2/C4 PPL by $0.4873/1.3771$; at length 2048, by $1.0570/2.6553$. The data support retaining one-shot QK-WANDA for the main method. They do not identify whether irreversible mask allocation, omitted residual terms, or other interactions cause the degradation, and do not rule out other adaptive algorithms.

Source paths, SHA256 hashes, and all 16 rows are retained in the paper's `results/iterative_archive.json`. The exporter checks settings, step coverage, and final sparsity. The single-seed ablation was evaluated separately from the three-seed comparisons.

K Relation to Published QK Reconstruction Objectives

K.1 From a reconstruction quadratic to scalar deletion

Consider one query head and its associated key head, with the head labels suppressed. The matrix ΔW^Q is a change to that query projection, $\delta w = \mathrm{vec}(\Delta W^Q)$ stacks its columns, and G_U is the unmasked quadratic coefficient from Equation (33). With keys fixed, the loss is

$$\begin{aligned}\mathcal{L}(\Delta W^Q) &= \frac{1}{N}\sum_{s=1}^{N}\left\|(\Delta W^Q X^{(s)})^\top K^{(s)}\right\|_F^2 \\ &= \frac{1}{N}\sum_{s=1}^{N}\left\|K^{(s)\top}\Delta W^Q X^{(s)}\right\|_F^2 = \delta w^\top G_U \delta w.\end{aligned} \tag{53}$$

Up to a common positive normalization factor, AESPA gives the corresponding Q/K reconstruction targets and vectorized coefficient in Equations (15)–(16) and (19) (Kim et al., 2024). Its key orientation differs from ours; transposing the product preserves its Frobenius norm. Evaluating this published type of quadratic at a scalar deletion, with vectorized index p as in Section A.5, gives our algebraic specialization:

$$\delta w = -w_p e_p \quad \Longrightarrow \quad \mathcal{L} = w_p^2 [G_U]_{pp}. \tag{54}$$

This evaluates the Q/K reconstruction quadratic for deletion of an original scalar weight. AESPA also introduces a product-of-expectations approximation in its Equation (20); our sequence-local accumulators retain the sum of products in Equation (33).

BOA's Equations (11)–(12) use related Q/K Hessians, and (14)–(15) incorporate RoPE through averaged rotated factors (Kim et al., 2025). That establishes a prior rotary-aware quantization construction. Our QK-WANDA-R proof evaluates the exact finite change of an original scalar coordinate and retains its token-specific rotation, input square, and causal pair restriction. QK-WANDA uses the corresponding causal restriction without rotations. The resulting statistics retain the causal, token-specific interactions of the scalar perturbation.

K.2 A channel deletion is a different perturbation

For one sequence and one unmasked head, suppress sequence/head labels. Using the same unscaled convention, decompose the product as

$$B = Q^\top K = \sum_i Q_{i:}^\top K_{i:}. \tag{55}$$

Deleting a complete matched Q/K dimension i removes the summand $Q_{i:}^\top K_{i:}$. Its uncompensated energy is $\|Q_{i:}\|_2^2 \, \|K_{i:}\|_2^2$. CORP uses an expected channel-contribution energy to rank such dimensions, then reconstructs removed contributions through closed-form compensation (Zhang & Yang, 2026, Sections 3.2–3.4). In our weight layout, this changes whole matched output rows of Q and K. Deleting just W_{ij}^Q instead produces the perturbation $-W_{ij}^Q X_{j:}^\top K_{i:}$, with a different score for each input coordinate j. The difference is the pruning unit, not the use of the word "row" or "column" under transposed conventions.

The contribution is the causal, compensation-free elementwise pruning specialization, its exact GQA-aware scalar statistics, and its controlled evaluation. The cited prior methods do not instantiate this full procedure. The quadratic connection above identifies which underlying ideas are shared and which compression decision and accumulators are studied here.

L COST OF COMPUTING THE Q/K SCORES

Computational cost with multiple heads. Using the notation of Section A.1, let $D_Q = H_Q d_h$ and $D_K = H_K d_h$ be the total query and key projection widths. For N calibration sequences of a common length T, the statistic scans cost $O(NT(d + D_Q + D_K))$ and the grouped interaction products cost $O(NTD_K d)$, followed by elementwise score formation and sorting. The query-side interaction factors are reused by query heads sharing a key head; the key-side factors sum contributions from those query heads. Setting $N = H_Q = H_K = 1$ and $d_h = m$ gives the single-sequence, single-head costs in the main text.

Measurement scope. We distinguish score computation from the shared cost of running the model on calibration inputs. *Score only* starts with cached GPU-resident inputs X and raw projections Q, K, initializes fresh accumulators, processes all calibration sequences, and materializes both complete weight-score matrices. WANDA accumulates input energies separately for its Q and K modules; QK-WANDA accumulates causal opposite-projection statistics. The former never computes QK statistics in this benchmark. *Calibration + score* additionally includes a decoder-block forward pass for each sequence, collecting the respective statistics through the production hooks. Both measurements exclude checkpoint loading, tokenization, CPU/GPU transfers, mask sorting and application, and all evaluation diagnostics.

Controlled protocol. We load Llama-3.2-1B/3B and Llama-3.1-8B/70B in FP16 and use 256×32 C4 calibration tokens with seed 0 for this timing study. Both methods receive identical dense hidden states at every layer. We retain all checkpoint weights on CPU and move one decoder block at a time to the same NVIDIA A100 PCIe 40GB GPU. Activations are captured once outside the measured intervals. Every measured call constructs fresh statistics; sequence batch size is one, accumulation uses FP32, and TF32 is disabled. Each method has two warm-up calls followed by five timed repetitions per block, with alternating method order. CUDA synchronization brackets each wall-clock measurement. The reported times sum measurements over every actual decoder block before taking the median over repetitions; they are not extrapolated from one layer. Cached replay and hooked forward passes produce exactly matching score tensors for each method.

Table 21: Cost of scoring all Q/K weights with 256×32 calibration tokens, in seconds, on one A100 40GB. Loading, data movement, mask construction, and evaluation are excluded. Medians over five timing repetitions; each ratio is QK-WANDA/WANDA. The right-hand group includes calibration forward passes, but neither group measures total pruning time.

	Score only			Calibration + score		
Model	WANDA	QK-WANDA	Ratio	WANDA	QK-WANDA	Ratio
1B	0.356	1.579	4.43×	4.111	5.374	1.31×
3B	0.665	2.973	4.47×	7.628	10.007	1.31×
8B	1.037	4.270	4.12×	10.710	14.015	1.31×
70B	1.880	9.064	4.82×	33.337	45.536	1.37×

In this 256×32-token benchmark, including calibration forward passes gives a QK-WANDA/WANDA runtime ratio of 1.31–1.37. Isolated score construction takes 4.12–4.82 times as long; the smaller combined ratio reflects the shared forward computation. These intervals exclude data movement and mask construction.

Score construction precedes mask selection, so these timings do not depend on a chosen sparsity level. They describe the current implementation and calibration length; a different sequence length, batching strategy, or hardware can change the ratios. Five repetitions quantify timing variability, not variation over calibration datasets. Per-layer wall times, CUDA-event timings, checkpoint revisions, token hashes, and code hashes are retained in `output/score_cost_20260916/`; the portable table data and all five aggregate repetitions are in `results/score_cost.json`.

M WANDA WITH MATRIX-WIDE BUDGETS

We tested Wanda with a separate budget for each full Q and K matrix on the three 70B/72B checkpoints. Every projection matched the corresponding separate QK-Wanda run's integer deletion count. Checkpoints, tokenizers, seed 0, 524,288 calibration tokens segmented into length-64 sequences, and PPL context 2048 were fixed. H200 runs used FP16 for Llama and BF16 for Qwen. Biases and non-Q/K weights remained dense.

Pruning was sequential across blocks: sampled tokens matched, while later activations could differ because of earlier masks. QK error used common dense inputs and the original 16 held-out C4 sequences of length 512; each pass reproduced the published 50% shared-mask anchor. Table 22 compares the scores under identical budgets. QK-Wanda produced lower measured WikiText-2/C4 PPL in nine/nine of nine configurations on Llama 2, seven/four on Llama 3, and five/seven on Qwen; QK error was lower in all 27. At 80% on Qwen, matrix-Wanda had lower PPL than separate QK-Wanda despite larger QK error; the C4 difference is hidden by rounding.

Table 22: Scoring at matched matrix budgets: Wanda and causal pre-RoPE QK-Wanda with separate Q/K budgets and identical per-projection deletion counts. Lower is better. Bold marks lower displayed values within each pair, including ties. Errors are block-mean relative squared QK errors (%).

	WikiText-2 PPL		C4 PPL		Rel. QK error (%)	
Sparsity	Wanda matrix	QK-Wanda separate	Wanda matrix	QK-Wanda separate	Wanda matrix	QK-Wanda separate
			Llama-2-70B			
10%	**3.32**	**3.32**	**5.71**	**5.71**	**0.00**	**0.00**
20%	**3.32**	**3.32**	**5.71**	**5.71**	0.01	**0.00**
30%	**3.32**	**3.32**	**5.71**	**5.71**	0.05	**0.01**
40%	3.33	**3.32**	5.72	**5.71**	0.18	**0.03**
50%	3.36	**3.34**	5.74	**5.72**	0.56	**0.11**
60%	3.43	**3.37**	5.81	**5.75**	1.46	**0.29**
70%	3.61	**3.48**	5.99	**5.84**	3.39	**0.71**
80%	4.13	**3.84**	6.53	**6.14**	7.42	**1.67**
90%	5.76	**5.62**	8.76	**7.96**	16.21	**4.27**
			Llama-3.1-70B			
10%	**2.81**	**2.81**	**7.11**	**7.11**	**0.00**	**0.00**
20%	**2.81**	**2.81**	7.12	**7.11**	0.03	**0.00**
30%	2.86	**2.82**	7.15	**7.13**	0.17	**0.03**
40%	3.00	**2.88**	7.28	**7.19**	0.63	**0.12**
50%	3.35	**3.06**	7.61	**7.41**	1.58	**0.38**
60%	4.14	**3.95**	**8.46**	9.33	3.14	**1.06**
70%	**6.25**	8.89	**10.76**	16.90	5.36	**2.59**
80%	**17.69**	26.00	**24.21**	43.09	8.52	**5.50**
90%	208.08	**97.00**	**120.85**	138.71	13.63	**10.18**
			Qwen2.5-72B			
10%	**3.88**	**3.88**	**9.26**	**9.26**	**0.00**	**0.00**
20%	**3.88**	**3.88**	**9.26**	**9.26**	0.01	**0.00**
30%	3.89	**3.88**	**9.26**	**9.26**	0.05	**0.02**
40%	3.90	**3.89**	9.27	**9.26**	0.13	**0.04**
50%	3.92	**3.90**	9.29	**9.28**	0.28	**0.09**
60%	3.97	**3.95**	9.33	**9.31**	0.54	**0.18**
70%	**4.07**	**4.07**	9.42	**9.40**	0.96	**0.35**
80%	**4.31**	4.45	**9.64**	**9.64**	1.64	**0.68**
90%	**5.25**	5.61	**10.46**	10.47	2.81	**1.41**

M.1 ZERO-SHOT RESULTS

Table 23 reports the seven full task splits using the exact saved masks at 30%, 50%, and 80%. Evaluation documents and prompts matched the original comparison. Relative to row-wise Wanda, the matrix control improved the mean in three of nine configurations and worsened it in six. Shared QK-Wanda improved over the matrix control in six of nine, but lost on Llama 2 70B at 50% and Qwen2.5-72B at 30%/50%. At 80% on Llama-3.1-70B, the matrix control had lower PPL than shared QK-Wanda but much lower mean accuracy (57.74% versus 66.45%). The separate QK allocation was not evaluated on these tasks in this control, so the accuracy comparison does not isolate the scoring rule. All results use one calibration seed, and small differences do not establish statistical significance.

Table 23: Zero-shot accuracy (%) for the matrix-budget control and the original comparison. Wanda uses row budgets; Wanda (matrix) ranks weights within each projection; QK-Wanda uses shared Q/K budgets. Each section reuses the same dense reference and evaluation examples. Bold marks the highest displayed sparse score at each sparsity, including ties; Avg. is the unweighted seven-task mean.

Method	Sparsity	BoolQ	RTE	HellaSwag	WinoGrande	ARC-Easy	ARC-Challenge	OpenBook QA	Avg.
Llama-2-70B									
Dense	0%	83.76	67.87	83.81	77.98	80.98	57.34	48.80	71.50
WANDA	30%	83.43	**68.95**	83.72	77.82	80.72	**57.42**	48.40	71.50
WANDA (matrix)	30%	**83.70**	68.23	83.76	**78.06**	80.89	57.17	**48.60**	71.49
QK-WANDA	30%	83.64	**68.95**	**83.79**	77.90	**81.02**	**57.42**	**48.60**	**71.62**
WANDA	50%	82.63	69.31	83.38	**77.98**	80.89	57.00	48.40	71.37
WANDA (matrix)	50%	83.49	**71.12**	83.36	77.66	**81.06**	**57.51**	48.40	**71.80**
QK-WANDA	50%	**83.94**	68.95	**83.79**	77.90	81.02	57.17	**48.60**	71.62
WANDA	80%	76.51	65.70	76.66	68.03	78.11	50.34	45.60	65.85
WANDA (matrix)	80%	80.92	**73.29**	78.52	**78.06**	78.75	53.16	48.00	70.10
QK-WANDA	80%	**84.04**	70.40	**82.54**	77.27	**81.44**	**56.66**	**50.00**	**71.76**
Llama-3.1-70B									
Dense	0%	85.44	69.68	85.03	79.95	86.66	64.93	47.80	74.21
WANDA	30%	85.35	70.04	**84.86**	79.56	86.36	**64.76**	**48.20**	74.16
WANDA (matrix)	30%	**85.50**	70.40	84.60	79.40	86.41	64.16	47.20	73.95
QK-WANDA	30%	85.26	**70.76**	84.85	**79.64**	**86.62**	64.33	47.80	**74.18**
WANDA	50%	84.74	**70.04**	84.32	79.01	85.94	62.97	**48.20**	73.60
WANDA (matrix)	50%	82.75	69.68	82.56	79.08	84.64	61.26	47.80	72.54
QK-WANDA	50%	**85.11**	**70.04**	**84.54**	**79.24**	**86.20**	**63.82**	47.60	**73.79**
WANDA	80%	76.06	68.59	**76.62**	**69.14**	80.89	54.01	**41.20**	**66.64**
WANDA (matrix)	80%	71.99	57.76	67.76	59.27	69.82	46.16	31.40	57.74
QK-WANDA	80%	**77.80**	**71.12**	73.85	64.88	**81.61**	**55.72**	40.20	66.45
Qwen2.5-72B									
Dense	0%	89.17	77.26	86.06	78.14	83.38	62.63	46.60	74.75
WANDA	30%	89.27	77.26	86.13	77.51	**83.33**	62.46	46.80	74.68
WANDA (matrix)	30%	**89.33**	**77.98**	**86.16**	**78.45**	83.25	**62.54**	**47.40**	**75.02**
QK-WANDA	30%	89.17	77.26	86.04	77.58	83.16	62.46	46.60	74.61
WANDA	50%	89.11	**76.90**	85.90	**77.82**	83.42	**62.12**	**47.80**	**74.72**
WANDA (matrix)	50%	89.24	**76.90**	**86.12**	77.66	83.12	61.86	47.20	74.59
QK-WANDA	50%	**89.42**	76.53	85.96	77.58	**83.84**	61.86	46.60	74.54
WANDA	80%	87.55	**77.26**	83.90	**76.40**	83.25	60.15	45.60	73.44
WANDA (matrix)	80%	88.65	73.65	84.73	75.69	83.00	60.84	**46.00**	73.22
QK-WANDA	80%	**89.27**	75.09	**85.06**	76.24	**85.23**	**62.29**	44.60	**73.97**