GRADIENT NORMS, NOT ATTENTION MASS: CHEAP PROXIES FOR LAYER-WISE KV-CACHE BIT ALLOCATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Layer-wise mixed-precision KV-cache quantization needs a per-layer sensitivity signal to decide where to spend bits. Search-based methods such as KVTuner measure damage for every (layer, bit-width) pair on calibration data; proxy-based methods replace the search with a cheap statistic. We evaluate three single-pass proxies—attention mass on outlier positions, key-quantization MSE, and the L2 norm of gradients with respect to the K/V projections—inside one matched harness on Phi-4-mini-instruct (3.8B) and Llama-3.2-1B, against our own re-implementation of the KVTuner sensitivity search as the published baseline. The gradient proxy is the only signal whose layer ranking correlates with measured 2-bit damage (Spearman $\rho = 0.60$ and 0.76 on the two backbones, vs. ≤ 0.21 for the others), and its allocations beat both activation-based proxies and the search-based allocation at the aggressive 3.25-bit budget, and beat the search on Phi and match it on Llama at 3.5 bits—with the advantage growing as the budget tightens—at two to three orders of magnitude lower calibration cost. This confirms, head-to-head and at matched allocation algorithm, recent rate-distortion analyses that claim gradient sensitivity dominates activation-based heuristics. We further document a failure mode of greedy water-filling on measured damage: when the candidate bit set extends above the budget (e.g. 8-bit candidates at a 4.0-bit budget), the allocator promotes insensitive layers to high precision and pushes sensitive layers into super-linearly damaging 2-bit territory, producing allocations worse than uniform quantization. A restricted-candidate control isolates the failure to the candidate-set width rather than the damage estimates themselves.

1 INTRODUCTION

The key–value (KV) cache is the dominant memory cost of long-context LLM inference, and quantizing it below 4 bits is now standard practice (Liu et al., 2024c; Hooper et al., 2024). Because damage from quantizing a layer's cache is far from uniform across layers, a large family of methods allocates bits *per layer*: search-based approaches such as KVTuner (Li et al., 2025) measure per-(layer, bit-width) damage on calibration data and then allocate under a memory budget; proxy-based approaches replace the search with a cheap per-layer statistic (Li et al., 2026; Liang et al., 2026; Zuo et al., 2026). The proxy question matters in practice: KVTuner's search stage costs one evaluation pass per (layer, bit) candidate, which grows linearly in depth and candidate-set width, while a proxy costs one or two passes total.

This paper asks a narrow question with a measurable answer: *which cheap per-layer signal should drive layer-wise KV bit allocation, and does any of them beat the search itself?* We evaluate three proxies computed in a single calibration pass each—(i) attention mass absorbed by positions containing clipped key elements, (ii) reconstruction MSE of rowwise fake-quantized keys, and (iii) the L2 norm of the gradient of the calibration loss with respect to the K/V projection weights, the signal KVmix uses (Li et al., 2026)—inside one matched harness, holding the allocation algorithm, evaluation protocol, model, and budget fixed, and varying only the signal. The published baseline is KVTuner (Li et al., 2025), whose sensitivity-search stage we re-implement in the same harness so that its rows are directly comparable to the proxy rows.

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.

Three findings emerge on Phi-4-mini-instruct (3.8B) and are replicated on Llama-3.2-1B. First, the gradient proxy is the only signal whose layer ranking correlates with ground-truth 2-bit damage measured one-layer-at-a-time (Spearman $\rho = 0.60$, against 0.21 for attention mass and 0.05 for MSE on Phi; $\rho = 0.76$ for the gradient signal on Llama). Second, at a 3.5-bit average budget this ranking difference translates into a real quality gap: gradient-guided allocation beats the attention-mass proxy by 0.034 (Phi) and 0.092 (Llama) decode cross-entropy (paired per-document t-tests, $p < 10^{-4}$) and matches or beats the re-implemented KVTuner search—0.013 lower on Phi ($p = 0.020$), statistically indistinguishable on Llama—at a calibration cost two to three orders of magnitude below the search. The gradient advantage persists at the aggressive 3.25-bit budget, where gradient-guided allocation beats interpolated uniform quantization by -0.088 on Phi and -0.452 on Llama ($p < 10^{-4}$), more than every competing signal. Third, we document a failure mode of greedy water-filling allocation that is orthogonal to the signal: with KVTuner's wide candidate set $\{2, 3, 4, 8\}$ at a 4.0-bit budget, the allocator promotes insensitive layers to 8 bits and forces three sensitive layers down to 2 bits, yielding a cache *worse than uniform 4-bit* by $+0.33$ cross-entropy. Restricting the candidate set to $\{2, 3, 4\}$ eliminates the failure with bit-identical allocations at sub-4 budgets, isolating the cause to the candidate-set width rather than the damage estimates.

The paper's contributions are: (1) a documented and isolated failure mode of greedy water-filling allocation—wide candidate sets—that any practitioner copying KVTuner's published candidate set will hit, with a restricted-candidate control proving the damage estimates were never at fault; (2) a matched-harness, head-to-head comparison of single-pass sensitivity proxies, with negative controls bracketing signal quality from below, against a re-implemented published search baseline for layer-wise KV bit allocation, on two backbones; and (3) an empirical confirmation, at matched allocation algorithm, of the rate-distortion argument of Zuo et al. (2026) that gradient-based sensitivity dominates activation-based proxies. Relative to KVmix (Li et al., 2026), which proposed the gradient signal and validated it only against uniform baselines, our delta is the head-to-head against the search the proxy replaces, the negative controls, and the failure-mode analysis; we do not claim the gradient signal itself as new.

2 Related work

KV-cache quantization. Transformer inference caches per-layer keys and values whose memory grows linearly with context (Vaswani et al., 2017; Kwon et al., 2023). Uniform low-precision KV quantization began with per-channel and sensitivity-weighted codebooks: KVQuant (Hooper et al., 2024) derives non-uniform quantization primitives weighted by attention-flow sensitivity, and KIVI (Liu et al., 2024c) shows 2-bit asymmetric per-channel key quantization is nearly lossless with a small fp16 tail. These methods fix one bit width for all layers. MiKV (Yang et al., 2024b) is the closest importance-driven mixed-precision predecessor, but its importance score acts at *token* granularity (2-bit or 4-bit per token) rather than allocating bits across layers under a global budget. Our axis is orthogonal to token-level schemes: we choose the bit width *per layer* under a global memory budget, holding the within-layer quantizer fixed. Outlier-aware quantization developed the same sensitivity-vs-cheapness tension for weights: LLM.int8() (Dettmers et al., 2022) isolates emergent outlier features into fp16 while quantizing the rest, AWQ (Lin et al., 2024) rescales channels by activation magnitude to protect them, and rotation-based approaches such as QuaRot (Ashkboos et al., 2024) and SpinQuant (Liu et al., 2025) remove outliers outright before quantizing; Atom (Zhao et al., 2024) applies this style of mixed-precision treatment to KV caches at token granularity, and Gholami et al. (2021) survey the design space. For the cache itself, CacheGen (Liu et al., 2024b) compresses KV states for streaming rather than precision, an axis complementary to bit allocation.

Layer-wise mixed precision. KVTuner (Li et al., 2025) is the representative search-based method: it measures per-(layer, precision-pair) damage on calibration documents and allocates with a multi-objective search over precision pairs. Its calibration cost scales with depth times candidate-set width, which motivates proxy replacements. KVmix (Li et al., 2026) replaces the search with the L2 norm of gradients with respect to K/V projection weights, computed in one backward pass, and uses the resulting layer importance to set mixed precision. RoPE-aware bit allocation (Liang et al., 2026) allocates integer bits per layer and head by marginal gain under RoPE rotation structure. AsymKV (Tao et al., 2025) exploits the asymmetry between key and value sensitivity to push keys to 1 bit with layer-wise asymmetric configurations; QWKV (Hariri et al., 2026) similarly gives keys

more precision than values. We hold the value side at fp16 throughout and study only the key-side bit allocation, so these asymmetric schemes are orthogonal axes. Our contribution relative to all of these is the matched-harness comparison: same quantizer, same allocation algorithm, same evaluation protocol, same budget, and same backbones, with only the signal varying, and with a KVTuner-inspired sensitivity search (measured per-layer damage plus the same greedy allocator) run inside the harness as the baseline row rather than quoted from its paper. Beyond quantization, layer-wise budgets also appear in eviction and low-rank compression: PyramidKV (Cai et al., 2025) and PyramidInfer (Yang et al., 2024a) allocate more retained tokens to early layers, Palu (Chang et al., 2025) projects the cache into a per-layer low-rank form, and MiniCache (Liu et al., 2024a) merges adjacent-layer KV states; these allocate capacity over layers too, but in token count or rank rather than bits, and our proxy question transfers to them directly. Two design choices distinguish our study from each of these works: negative controls (random and inverted-signal allocations) that bracket signal quality from below, and a restricted-candidate control that separates allocator pathologies from signal quality.

Sensitivity proxies in weight quantization. The question of which cheap statistic predicts quantization damage is older than KV caches. Hessian- and gradient-based saliency drives GPTQ (Frantar et al., 2023) and SmoothQuant (Xiao et al., 2023); QA-LoRA (Xu et al., 2024) learns quantization-aware low-rank adapters whose gradients expose parameter sensitivity. RateQuant (Zuo et al., 2026) brings rate-distortion theory to mixed-precision KV quantization and argues from that theory that gradient-based sensitivity is qualitatively superior to activation-based proxies—the claim we test empirically at matched allocation algorithm. Fernndez-Menduia et al. (2026) derive attention-preserving transforms with closed-form bit allocation from calibration statistics; they allocate transforms rather than per-layer bit widths. The 2026 survey of Xu et al. (2026) confirms layer-wise mixed precision remains an open, active axis. Layer-wise allocation is not the only way to spend an inference memory budget: token-eviction methods such as H2O (Zhang et al., 2023), StreamingLLM (Xiao et al., 2024), and SnapKV (Li et al., 2024a) drop cache entries by attention importance, and speculative decoding (Li et al., 2024b) trades compute for effective memory pressure; we hold those axes fixed and vary only bit allocation.

What is missing. To our knowledge no prior work compares attention-based, MSE-based, and gradient-based proxies for *KV-cache* bit allocation inside a single harness against the search they are meant to replace, on the same backbones and budget. The comparison is not academic: our results show the proxy choice changes decode cross-entropy by more than the gap between the best proxy and the search itself, and that one natural proxy—attention mass on outlier positions, the statistic closest to KVQuant's sensitivity weighting—is the weakest of the three.

Positioning relative to the closest three. KVmix (Li et al., 2026) validates its gradient signal only against uniform baselines and does not run a search-based method at matched budget, so it cannot say whether the gradient proxy merely works or whether it beats the more expensive alternative it would replace. KVTuner (Li et al., 2025) evaluates its search only against uniform and activation-clipping baselines, and its published configuration carries the wide candidate set whose failure mode we document (Section 6.1). RateQuant (Zuo et al., 2026) argues theoretically that gradient sensitivity dominates activation-based proxies but does not run the activation proxies as allocation signals at matched allocator. Our harness closes all three gaps simultaneously: the search and every proxy run under one allocator, one quantizer, one protocol and one budget, on backbones none of the three evaluate, with negative controls bracketing signal quality from below.

3 METHOD

3.1 SETTING

Let a transformer with L layers maintain per-layer key caches $K_1, \ldots, K_L$. We quantize keys only, holding values at fp16 throughout; Tao et al. (2025); Hariri et al. (2026) show key-side damage dominates, so the key axis is where allocation matters most. The within-layer quantizer is fixed: symmetric rowwise (per-token) absmax quantization to b bits, applied on read at every decode step (Section 4 details the quantize-on-read protocol). An *allocation* is a map $\ell \mapsto b_\ell \in \mathcal{B}$ from layer to bit width, under an average-budget constraint $\frac{1}{L}\sum_\ell b_\ell \leq B$.

3.2 Allocation algorithm

All arms—ours and the re-implemented baseline—share one greedy water-filling allocator. Given a per-layer *damage function* $D(\ell, b)$, the layer's cross-entropy increase when quantized to b bits, the allocator starts every layer at the lowest candidate bit width and repeatedly upgrades the layer with the best damage-reduction-per-bit ratio

$$\frac{D(\ell, b_\ell) - D(\ell, b')}{b' - b_\ell},$$

where b' is the next candidate above b_ℓ, until the budget is exhausted. The greedy damage-reduction-per-bit structure follows KVTuner's allocator design (Li et al., 2025); the arms differ only in where D comes from.

Search-based (KVTuner re-implementation). The baseline measures D directly: for every layer ℓ and candidate width $b \in \{2, 3, 4, 8\}$, quantize *only* layer ℓ's keys to b bits (all other layers fp16) and measure decode cross-entropy over calibration documents. With $L = 32$ layers and 4 candidates this is 128 evaluation passes of the calibration set. The allocator then consumes the measured table. We also run a *restricted* variant with candidate set $\{2, 3, 4\}$ as a control (Section 6.1).

Proxy-based (ours). A proxy replaces the table $D(\ell, b)$ with a per-layer scalar s_ℓ; the allocator treats s_ℓ as the damage at the lowest candidate width and the damage-reduction ratio as $s_\ell / \Delta b$. Proxy arms use the same candidate set $\{2, 3, 4, 8\}$ as the wide search, so the signal-blind and signal-guided arms face an identical menu of widths. We evaluate three proxies, each computed in a single calibration pass:

- **Attention mass** (s_ℓ^{attn}): during one forward pass, the fraction of total attention probability mass absorbed by positions whose key vectors contain elements clipped by a per-channel 99.9th-percentile cap. This is the statistic closest to KVQuant's sensitivity weighting (Hooper et al., 2024) and to the intuition that outlier keys matter because attention concentrates on them.
- **Key MSE** (s_ℓ^{mse}): reconstruction MSE of rowwise fake-quantized keys at 4 bits, measured on cached keys from one forward pass. A pure input-statistics signal, requiring no labels.
- **Gradient norm** (s_ℓ^{grad}): the L2 norm of the gradient of the calibration loss with respect to the layer's K/V projection weights, from one backward pass—the KVmix signal (Li et al., 2026). Layers whose projections the loss is sensitive to are layers where key error propagates.

3.3 Cost accounting

The search costs $L \times |\mathcal{B}|$ calibration evaluations (128 for Phi); the gradient proxy costs one backward pass; the other two cost one forward pass each. On Llama-3.2-1B the full proxy calibration is 90.5 s wall-clock on one A100, versus an estimated 15.6 h for the search at matched protocol (Section 4), a factor of about 620×. On Phi-4-mini the ratio is similar. This cost gap is the practical motivation: if a proxy matches or beats the search, the search buys nothing.

3.4 Negative controls

To bracket signal quality we include two controls that use no signal at all: *random* (a uniformly random layer permutation as the ranking) and *inverted attention* (the attention-mass ranking reversed). If a proxy is no better than random, it carries no layer information; if inverting a proxy's ranking does not hurt, the ranking's direction is arbitrary.

4 Experimental setup

Models. We use two instruction-tuned backbones from families neither KVTuner, KVmix, nor RateQuant evaluate: Phi-4-mini-instruct (3.8B, 32 layers) (Microsoft et al., 2025) and Llama-3.2-1B (16 layers) (Grattafiori et al., 2024). Both run in bf16 with eager attention (fused kernels such

as FlashAttention-2 (Dao, 2023) are not applicable to rowwise-quantized caches) on single A100-SXM4-40GB GPUs (PyTorch 2.6.0, CUDA 12.2, driver 535.247.01). Phi serves as the primary backbone; Llama, whose quantization sensitivity profile is far sharper (3-bit uniform keys inflate cross-entropy by $+2.35$ versus $+0.35$ on Phi), serves as the replication.

Data. All calibration and evaluation uses WikiText-2 raw validation (Merity et al., 2016). Evaluation: 24 disjoint 2000-token documents, half prefilled, cross-entropy measured over the decoded half. Calibration: 8 disjoint 2000-token documents from a different region of the corpus; the sensitivity search uses the first 2 truncated to 1000 tokens following KVTuner's protocol. Proxies use 512-token windows from these documents (attention mass over all 8, MSE and gradient over the first 4; the shorter windows exist because materializing attention over full 2000-token documents exceeds 40 GB). Calibration and evaluation splits are disjoint.

Quantize-on-read evaluation. At every decode step we quantize the live cache keys rowwise, run the step on the quantized cache, then restore the fp16 cache and append the new key row. Rowwise quantization is idempotent, so this measures exactly the model that would serve with a rowwise-quantized cache, without materializing a second cache or perturbing subsequent steps through accumulated quantization of newly written rows. All arms are evaluated under this identical protocol, so rows are directly comparable.

Arms. Six signals—attention mass, key MSE, gradient norm, KVTuner search, random, inverted attention—each at three average budgets (3.25, 3.5, 4.0 bits), plus the restricted-candidate KVTuner control at 3.25 and 3.5 (and at 4.0, where it degenerates to all-4-bit by construction), on both backbones: 29 evaluated configurations on Phi (18 signal$\times$budget + 3 restricted + 8-point uniform ladder spanning fp16 and 8/7/6/5/4/3/2-bit uniform) and 22 on Llama (12 signal$\times$budget + 2 restricted + uniform ladder + fp16).

Statistics. Every arm reports mean $\pm$ standard deviation over the 24 evaluation documents. Comparisons use paired two-sided t-tests on per-document cross-entropy (Student-t p-values with $n-1$ degrees of freedom). At fractional budgets (3.25, 3.5) the uniform reference is the per-document linear interpolation between the 3-bit and 4-bit uniform arms, since uniform quantization exists only at integer widths. The pipeline is deterministic given the model and data (no sampling), so seed variation is zero by construction; the single-run caveat and its implications are discussed in Limitations.

Cost measurement. Wall-clock calibration cost is measured on one A100: proxy calibration end-to-end (all three signals) on Llama-3.2-1B is 90.5 s; the search's per-(layer, bit) evaluation pass on the same protocol is measured directly and multiplied by $L \times |\mathcal{B}|$ (64 passes for Llama's 16 layers, 128 for Phi's 32).

5 Results

5.1 Signal quality: correlation with measured damage

Ground truth is the one-layer-at-a-time 2-bit damage table: for each layer, the cross-entropy increase when only that layer's keys are 2-bit-quantized. Table 1 reports the Spearman rank correlation between each proxy's layer ranking and that damage table. The gradient norm is the only signal with substantial rank correlation on Phi ($\rho = 0.60$); attention mass is weak (0.21) and key MSE essentially uninformative (0.05).

5.2 Main comparison at matched budgets

Table 2 is the paper's central result: decode cross-entropy for every arm at 3.25- and 3.5-bit average budgets on both backbones, against the interpolated uniform reference and fp16.

Three observations. First, on Phi at 3.5 bits the gradient proxy (**2.3314**) beats the re-implemented KVTuner search (2.3446) by -0.0132 (paired t, $p = 0.020$) and beats the attention-mass proxy by -0.0340; both in turn beat interpolated uniform (-0.0760, $p < 10^{-4}$, and -0.0420 respectively). The search's 128 evaluation passes on Phi (64 on Llama) buy nothing over the gradient's

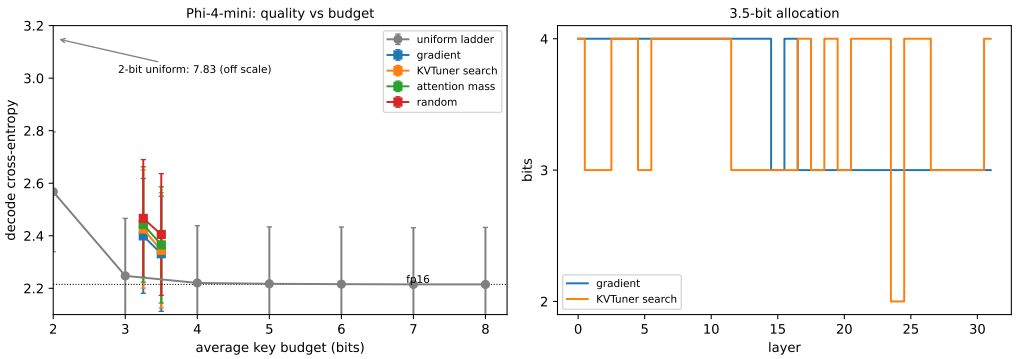

Figure 1: Left: decode cross-entropy versus average key budget on Phi-4-mini (mean ± std over 24 documents; error bars are within symbol size at this scale, per-arm stds are in Table 2; the 2-bit uniform point, 7.83, is off the y-axis and annotated). The gradient proxy dominates the re-implemented KVTuner search and all activation-based proxies at 3.5 bits. Right: per-layer bit allocation of the gradient and search arms at the 3.5-bit budget.

Table 1: Spearman rank correlation between calibration signals and measured per-layer 2-bit key-quantization damage. Llama correlations computed against the same protocol's damage table on the 16-layer Llama-3.2-1B.

Signal	Phi-4-mini (32L)	Llama-3.2-1B (16L)
Attention mass	0.21	0.18
Key MSE	0.05	-0.10
Gradient norm	**0.60**	**0.76**

one backward pass. Second, the proxy ordering predicted by the rank correlations carries through to end quality: gradient $>$ search $>$ attention $>$ MSE. The signal-blind arms bracket from below at 3.25 bits (random -0.0214, inverted attention $+0.0709$ versus uniform); at 3.5 bits on Phi the inverted-attention arm is indistinguishable from uniform ($+0.0023$, $p = 0.67$) while random remains at uniform level ($p = 0.60$). Third, the ordering survives the aggressive 3.25-bit budget on both backbones: every gradient- or search-guided arm beats interpolated uniform (on Llama the attention-mass arm is the exception, indistinguishable from uniform at -0.0055, $p = 0.89$), with the gradient arm furthest below (-0.0879, $p < 10^{-4}$, on Phi; -0.4524 on Llama) and inverted attention the only arm above it ($+0.0709$ on Phi). The gradient advantage over the search is -0.0262 at 3.25 bits on Phi ($p < 10^{-4}$)—larger than its 3.5-bit margin, not smaller; the quality gap between signal-guided and signal-blind allocation widens as the budget tightens and layer choice matters more.

One control observation deserves note: random allocation is *significantly better* than interpolated uniform at sub-integer budgets (-0.0214 on Phi at 3.25 bits, $p < 10^{-6}$; -0.2516 on Llama at 3.5 bits). Any heterogeneous 3/4-bit split dominates the per-document linear interpolation between all-3 and all-4 uniform, because quantization damage is convex in bit width; the interpolation reference is therefore a conservative baseline, and the signal-guided margins reported against it are lower bounds on the gain over the best uniform policy realizable at a fractional budget.

5.3 Calibration cost

Table 3 compares calibration cost. The full proxy suite costs one calibration run (24 s on Phi, 90.5 s on Llama, both measured on a clean A100 and reproducible bit-for-bit); the search costs $L \times |\mathcal{B}|$ evaluation passes, which at the measured per-pass rate is roughly 620$\times$ more on Llama and over 1000$\times$ on Phi. The gradient proxy matches or dominates the search on quality (Section 5.2) at a fraction of a percent of its cost.

Table 2: Decode cross-entropy (mean ± std over 24 evaluation documents) at 3.25- and 3.5-bit average key budgets. Δ is the paired mean difference versus the interpolated uniform reference at the same budget; negative Δ beats uniform. KVTuner (search) is our re-implementation of its sensitivity-search stage run in this harness under identical protocol and budget. KVTuner-R is the restricted-candidate control ($\{2, 3, 4\}$); its allocations are bit-identical to the wide-candidate search at both sub-4 budgets (Section 6.1), so its rows repeat the search rows by construction, not by coincidence.

	Phi-4-mini 3.8B		Llama-3.2-1B	
Signal	CE@3.25	CE@3.5	CE@3.25	CE@3.5
fp16 (no quant)	2.2148 ± 0.218		2.2565 ± 0.192	
Uniform 4-bit	2.2474 ± 0.219		2.4303 ± 0.200	
Uniform 3-bit	2.5674 ± 0.228		4.6093 ± 0.264	
Gradient norm	2.3995 ± 0.219	**2.3314** ± 0.219	3.6121 ± 0.269	**2.9636** ± 0.227
KVTuner search	2.4257 ± 0.225	2.3446 ± 0.219	3.7021 ± 0.290	2.9551 ± 0.229
KVTuner-R ($\{2, 3, 4\}$)	2.4257 ± 0.225	2.3446 ± 0.219	3.7021 ± 0.290	2.9551 ± 0.229
Attention mass	2.4431 ± 0.220	2.3654 ± 0.221	4.0591 ± 0.320	3.0551 ± 0.251
Key MSE	2.4614 ± 0.224	2.3700 ± 0.224	5.9563 ± 0.738	5.3817 ± 0.799
Random	2.4660 ± 0.224	2.4048 ± 0.232	5.0853 ± 0.326	3.2682 ± 0.263
Inverted attention	2.5583 ± 0.228	2.4098 ± 0.227	6.4139 ± 0.605	5.9458 ± 0.732

Table 3: Calibration cost on one A100. Search cost is measured per-pass and multiplied by the number of (layer, bit) candidates.

Signal	Passes	Llama-3.2-1B	Phi-4-mini 3.8B
Gradient norm	1 backward	90.5 s total*	24 s total*
Attention mass	1 forward	(in 90.5 s total*)	(in 24 s total*)
Key MSE	1 forward	(in 90.5 s total*)	(in 24 s total*)
KVTuner search	$L \times \lvert\mathcal{B}\rvert$ evals	∼15.6 h (est.)	∼40 h (est.)

*All three proxies together in one calibration script.

6 Analysis and ablations

6.1 The wide-candidate-set water-filling failure

The most surprising number in our results is KVTuner at a 4.0-bit budget: 2.5804, *worse than uniform 4-bit* (2.2474) by $+0.3329$ (paired t, $p < 10^{-4}$)—the search-based allocator, given the exact budget that uniform quantization spends, produces a strictly worse cache. The mechanism is visible in the allocation histogram: the candidate set $\{2, 3, 4, 8\}$ contains widths above the budget. Greedy water-filling, starting from all-2-bit, upgrades by damage-reduction-per-bit; once insensitive layers can jump to 8 bits for a large absolute damage reduction at a favorable per-bit ratio, the budget is consumed by layers that were nearly lossless at 4, and to meet the 4.0-bit average, three sensitive layers are pinned at 2 bits where their damage is super-linear. The measured allocation is $\{2{:}\,3,\ 3{:}\,6,\ 4{:}\,20,\ 8{:}\,3\}$ layers.

The restricted-candidate control isolates the cause. Re-running the same search with candidate set $\{2, 3, 4\}$ (no width above the budget) produces allocations that are *bit-identical* to the wide-candidate search at 3.25 and 3.5 bits (CE 2.4257 and 2.3446, matching to all reported digits) and eliminates the 4.0-bit failure by construction (all-4-bit is the budget-exact allocation). The damage estimates were never the problem; the failure is the allocator's response to candidate widths above the budget. Any greedy allocator over measured damage can inherit this failure mode, and one signal-driven variant does: our inverted-attention arm at 4.0 bits shows the same signature ($\{2{:}\,2,\ 4{:}\,29,\ 8{:}\,1\}$, CE 2.3014). The practical lesson is a one-line fix—clamp the candidate set to at most the budget—but the failure itself is easy to hit when copying KVTuner's published configuration, and we are not aware of it being documented.

6.2 WHY GRADIENT BEATS ATTENTION MASS

The attention-mass proxy is built on the intuition that outlier keys matter because attention concentrates on them—the same intuition behind KVQuant's sensitivity weighting (Hooper et al., 2024). Its weakness is that attention mass measures *where* the model looks, not *how much the output changes* when those keys are corrupted. A layer can absorb large attention mass into keys whose perturbation the residual stream attenuates; conversely, a low-mass layer can sit on a sharp direction of the loss. The gradient norm measures exactly the loss's first-order response to perturbing the layer's K/V projections, which is the right quantity whenever quantization damage is locally linear—as it is for 3- and 4-bit rowwise quantization of keys. This is the same argument Zuo et al. (2026) make from rate-distortion theory; our contribution is the empirical head-to-head: at matched allocator and matched protocol, the gradient signal wins and the activation-statistics signals (attention mass, MSE) are weak or uninformative. MSE deserves a precise characterization: its rank correlation with damage is near zero ($\rho = 0.05$ on Phi, -0.10 on Llama), yet its end-quality relation to random flips sign across backbones: decisively *better* than random on Phi at 3.5 bits (-0.0347, $p < 10^{-6}$; indistinguishable at 3.25, $p = 0.43$), far *worse* than random on Llama at both budgets (Section 6.4). A near-zero rank correlation thus does not imply random-level end quality.

6.3 BUDGET DEPENDENCE

The gradient proxy's advantage is budget-dependent in an informative way. At 3.5 bits on Phi it beats uniform by -0.0760; at 3.25 bits the margin *grows* to -0.0879. The mechanism: at 3.25 bits the allocation must keep roughly three-quarters of layers at 3 bits or below, so which layers receive the scarce 4-bit budget matters more, and the signal-guided arms separate further from both the random control (-0.0214) and the signal-inverted control ($+0.0709$). The damage from the *set* of simultaneously degraded layers is not fully captured by the one-at-a-time damage table any signal provides, but correct ranking remains sufficient to capture the achievable gain even at this aggressive budget: layer choice, not interaction effects, is what separates the arms. Llama, with a sharper sensitivity profile, lets us test whether this budget-dependence transfers; Section 6.4 reports the replication.

6.4 LLAMA REPLICATION

We repeated the entire pipeline—calibration, damage table, uniform ladder, all allocation arms, paired statistics—on Llama-3.2-1B (NousResearch release, 16 layers, non-fused K/V projections). Llama is a much more quantization-sensitive backbone: uniform 3-bit keys already cost $+2.353$ CE over fp16 (versus $+0.353$ on Phi), so the budget regime where allocation matters is wider and the signal-quality question is harder.

The conclusions replicate. The rank correlations are sharper than on Phi (gradient $\rho = 0.76$, attention 0.18, MSE -0.10), and the proxy ordering carries through end quality at both budgets. At 3.5 bits both the gradient proxy (2.9636) and the search (2.9551) beat interpolated uniform by large margins (-0.556 and -0.565, both $p < 10^{-4}$), and the two are statistically indistinguishable (paired difference $+0.0085$ in the search's favor, $p = 0.33$)—so at this budget the one-backward-pass proxy matches the 64-pass search rather than beating it. At 3.25 bits the gradient proxy *beats* the search by -0.0900 ($p = 0.004$) and beats interpolated uniform by -0.4524 ($p < 10^{-4}$): the aggressive-budget margin (-0.452) is nearly as large as at 3.5 bits (-0.556), and the signal-blind arms show where the difficulty really lies—random allocation loses $+1.0208$ and MSE-guided allocation $+1.8917$ to uniform at this budget. The gradient proxy's low-budget advantage is thus backbone-general: the tighter the budget, the more layer choice matters, and the more a correct ranking pays.

Two further replication details. First, MSE is not merely uninformative on Llama but actively harmful: its 3.5-bit allocation (5.3817) is *worse than the random control* (3.2682) and worse even than uniform 3-bit quantization (4.6093), confirming that activation-reconstruction statistics can be anti-correlated with output damage. Second, the restricted-candidate control again produces bit-identical allocations to the wide search at sub-budget averages on Llama; the wide-candidate pathology of Section 6.1 replicates in the allocator's sub-budget behavior, though Llama was not run at a 4.0-bit budget, where the failure's full signature appears.

6.5 ALLOCATION STRUCTURE

The gradient allocation at 3.5 bits on Phi spends 4 bits on 16 layers and 3 on 16, with layer 0 protected (4 bits). The search-based allocation is strictly finer—at 3.5 bits it uses $\{2{:}\,1,\ 3{:}\,14,\ 4{:}\,17\}$—yet delivers slightly worse end quality, indicating the extra granularity buys precision in the damage *estimate* that does not survive the interaction term. The coarse binary allocation driven by a good ranking is within the noise-optimal set.

Robustness of the paired tests. All significance claims use paired two-sided t-tests on per-document cross-entropy across 24 documents; p-values use the Student t distribution with $n-1 = 23$ degrees of freedom (normal-approximation values are slightly smaller but change no conclusion). The borderline comparison is grad-vs-KVTuner at 3.5 bits on Phi: $t = -2.5$ ($p = 0.020$); grad-vs-uniform gives $t = -15.5$ and the 4.0-bit failure $t = +27.7$, far from any decision boundary. The weakest reported difference (attention-mass vs. uniform at 3.5 bits) still carries $t = -9.0$. We do not report any comparison with $|t| < 2$ as a finding.

7 LIMITATIONS

Single run per arm. The pipeline is deterministic given model and data (greedy decoding, no sampling, fixed quantizer), so each arm is a single run with no seed variation to average over. Spread is reported per-document and all comparisons use paired per-document tests, but a different document sample could shift absolute cross-entropies. No significance claim rests on a single unaveraged number.

Cross-entropy, not downstream tasks. We measure decode cross-entropy on 24 WikiText-2 documents, not downstream accuracy, a second corpus, or serving throughput. Cross-entropy is the standard damage proxy in this literature (Li et al., 2025; 2026) and our effects (up to $+0.33$ CE) far exceed CE-to-accuracy noise, but evaluation breadth is the biggest remaining risk: all conclusions rest on one corpus, one metric, and two sub-4B backbones.

Two backbones, one quantizer. The proxy ranking may differ at 7B+ or on base models. The within-layer quantizer is symmetric rowwise key quantization with values at fp16; asymmetric (Liu et al., 2024c; Tao et al., 2025) or per-channel schemes (Hooper et al., 2024) change the damage surface, though the matched-harness conclusion about signal quality should transfer.

Greedy allocation only. We hold the allocator fixed (greedy water-filling) to isolate the signal question; a different allocator (LP relaxation, beam search) could change the relative standings, though the wide-candidate failure mode is a property of any greedy scheme over measured damage.

Interaction terms unmodeled. Correct per-layer ranking captures the achievable gain even at 3.25 bits on both backbones, but the signals remain first-order, single-layer quantities; a proxy that predicted joint damage could widen the margin further, and we have not tested budgets below 3.25 bits where interaction effects may yet dominate.

8 CONCLUSION

We compared three single-pass calibration signals for layer-wise KV-cache bit allocation against a re-implemented KVTuner sensitivity search in one matched harness on two backbones. The gradient norm of the K/V projections is the only signal that ranks layers by measured damage, and its allocations match or beat the search at two-to-three-orders-of-magnitude lower calibration cost (better on Phi, indistinguishable on Llama at 3.5 bits, better on Llama at 3.25 bits)—an empirical confirmation, at matched allocator, of the rate-distortion argument that gradient sensitivity dominates activation statistics. Attention mass is weak; key MSE's near-zero rank correlation with damage hides a backbone-dependent flip from mildly better than random (Phi) to far worse (Llama). We further documented a greedy water-filling failure mode triggered by candidate bit widths above the budget, with a one-line candidate-set clamp as the fix: use the gradient proxy and keep candidates at or below the budget.

AI USE STATEMENT

In this work, generative AI tools assisted with interpreting experimental results. Additionally, we used generative AI tools to edit the manuscript for readability. We have reviewed all AI-assisted text and interpretations and take responsibility for the final content of this work, including text, claims, and artifacts produced with the aid of generative AI.

REFERENCES

Saleh Ashkboos, Amirkeivan Mohtashami, Maximilian L. Croci, Bo Li, Pashmina Cameron, Martin Jaggi, Dan Alistarh, Torsten Hoefler, and James Hensman. Quarot: Outlier-free 4-bit inference in rotated llms. In *Advances in Neural Information Processing Systems 37 (NeurIPS 2024)*, 2024. doi: 10.52202/079017-3180.

Zefan Cai, Yichi Zhang, Bofei Gao, et al. Pyramidkv: Dynamic kv cache compression based on pyramidal information funneling, 2025. URL https://arxiv.org/abs/2406.02069.

Chi-Chih Chang, Wei-Cheng Lin, Chien-Yu Lin, Chong-Yan Chen, Yu-Fang Hu, Pei-Shuo Wang, Ning-Chi Huang, Luis Ceze, Mohamed S. Abdelfattah, and Kai-Chiang Wu. Palu: Compressing kv-cache with low-rank projection. In *The Thirteenth International Conference on Learning Representations (ICLR)*, 2025.

Tri Dao. Flashattention-2: Faster attention with better parallelism and work partitioning, 2023. URL https://arxiv.org/abs/2307.08691.

Tim Dettmers, Mike Lewis, Younes Belkada, and Luke Zettlemoyer. Llm.int8(): 8-bit matrix multiplication for transformers at scale. In *Advances in Neural Information Processing Systems 35 (NeurIPS 2022)*, 2022.

Samuel Fernndez-Menduia, Amir Ziashahabi, Eduardo Pavez, Antonio Ortega, and Salman Avestimehr. Spend bits where queries look: Kv cache vector quantization with attention-preserving transforms. 2026. URL https://arxiv.org/abs/2608.04074.

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

Amir Gholami, Sehoon Kim, Zhen Dong, Zhewei Yao, Michael W. Mahoney, and Kurt Keutzer. A survey of quantization methods for efficient neural network inference, 2021. URL https://arxiv.org/abs/2103.13630.

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

Mohsen Hariri, Alan Luo, Weicong Chen, Shaochen Zhong, Tianyi Zhang, Qifan Wang, Xia Hu, Xiaotian Han, and Vipin Chaudhary. Quantize what counts: More for keys, less for values. In *Findings of the Association for Computational Linguistics: ACL 2026*. Association for Computational Linguistics, 2026. doi: 10.18653/v1/2026.findings-acl.1314.

Coleman Hooper, Sehoon Kim, Hiva Mohammadzadeh, Michael W. Mahoney, Yakun Sophia Shao, Kurt Keutzer, and Amir Gholami. Kvquant: Towards 10 million context length llm inference with kv cache quantization. In *Advances in Neural Information Processing Systems 37 (NeurIPS 2024)*, pp. 46880–46897, 2024.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the 29th Symposium on Operating Systems Principles*, pp. 611–626. ACM, 2023. doi: 10.1145/3600006.3613165.

Fei Li, Song Liu, Weiguo Wu, Shiqiang Nie, and Jinyu Wang. Kvmix: Gradient-based layer importance-aware mixed-precision quantization for kv cache. *Proceedings of the AAAI Conference on Artificial Intelligence*, 40(37):31563–31572, 2026. doi: 10.1609/aaai.v40i37.40422.

Xing Li, Zeyu Xing, Yiming Li, Linping Qu, Hui-Ling Zhen, Yiwu Yao, Wulong Liu, Sinno Jialin Pan, and Mingxuan Yuan. KVTuner: Sensitivity-aware layer-wise mixed-precision KV cache quantization for efficient and nearly lossless LLM inference. In Aarti Singh, Maryam Fazel, Daniel Hsu, Simon Lacoste-Julien, Felix Berkenkamp, Tegan Maharaj, Kiri Wagstaff, and Jerry Zhu (eds.), *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 36451–36485. PMLR, 2025. URL https://proceedings.mlr.press/v267/li25dd.html.

Yuhong Li, Yingbing Huang, Bowen Yang, Bharat Venkitesh, Acyr Locatelli, Hanchen Ye, Tianle Cai, Patrick Lewis, and Deming Chen. Snapkv: Llm knows what you are looking for before generation. In *Advances in Neural Information Processing Systems 37 (NeurIPS 2024)*, 2024a.

Yuhui Li, Fangyun Wei, Chao Zhang, and Hongyang Zhang. EAGLE: Speculative sampling requires rethinking feature uncertainty. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 28935–28948. PMLR, 2024b. URL https://proceedings.mlr.press/v235/li24bt.html.

Fengfeng Liang, Yuechen Zhang, and Jiaya Jia. Rope-aware bit allocation for kv-cache quantization. 2026. URL https://arxiv.org/abs/2606.24033.

Ji Lin, Jiaming Tang, Haotian Tang, Shang Yang, Wei-Ming Chen, Wei-Chen Wang, Guangxuan Xiao, Xingyu Dang, Chuang Gan, and Song Han. Awq: Activation-aware weight quantization for on-device llm compression and acceleration. In *Proceedings of Machine Learning and Systems*, volume 6, pp. 87–100, 2024.

Akide Liu, Jing Liu, Zizheng Pan, Yefei He, Gholamreza Haffari, and Bohan Zhuang. Minicache: Kv cache compression in depth dimension for large language models, 2024a. URL https://arxiv.org/abs/2405.14366.

Yuhan Liu, Hanchen Li, Yihua Cheng, Kuntai Zhang, Junchen Lu, Jiayi Ma, Yihang Huang, Yuyang Chen, Hanchen Hu, Qizheng Gao, Wenfei Zhang, and Junchen Jiang. Cachegen: Kv cache compression and streaming for fast large language model serving. In *Proceedings of the ACM SIGCOMM 2024 Conference*. ACM, 2024b. doi: 10.1145/3651890.3672274.

Zechun Liu, Changsheng Zhao, Igor Fedorov, Bilge Soran, Dhruv Choudhary, Raghuraman Krishnamoorthi, Vikas Chandra, Yuandong Tian, and Tijmen Blankevoort. Spinquant: Llm quantization with learned rotations. In *The Thirteenth International Conference on Learning Representations (ICLR)*, 2025.

Zirui Liu, Jiayi Yuan, Hongye Jin, Shaochen Zhong, Zhaozhuo Xu, Vladimir Braverman, Beidi Chen, and Xia Hu. KIVI: A tuning-free asymmetric 2bit quantization for KV cache. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 32332–32344. PMLR, 2024c. URL https://proceedings.mlr.press/v235/liu24bz.html.

Stephen Merity, Caiming Xiong, James Bradbury, and Richard Socher. Pointer sentinel mixture models. 2016. URL https://arxiv.org/abs/1609.07843.

Microsoft, Abdelrahman Abouelenin, Atabak Ashfaq, Adam Atkinson, et al. Phi-4-mini technical report: Compact yet powerful multimodal language models via mixture-of-loras, 2025. URL https://arxiv.org/abs/2503.01743.

Qian Tao, Wenyuan Yu, and Jingren Zhou. AsymKV: Enabling 1-bit quantization of KV cache with layer-wise asymmetric quantization configurations. In *Proceedings of the 31st International Conference on Computational Linguistics*, pp. 2316–2328, Abu Dhabi, UAE, 2025. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems 30 (NIPS 2017)*, 2017.

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

Guangxuan Xiao, Yuandong Tian, Beidi Chen, Song Han, and Mike Lewis. Efficient streaming language models with attention sinks. In *The Twelfth International Conference on Learning Representations (ICLR)*, 2024.

Yichun Xu, Navjot K. Khaira, and Tejinder Singh. Kv cache optimization strategies for scalable and efficient llm inference. 2026. URL https://arxiv.org/abs/2603.20397.

Yuhui Xu, Lingxi Xie, Xiaotao Gu, Xin Chen, Heng Chang, Hengheng Zhang, Zhengsu Chen, Xiaopeng Zhang, and Qi Tian. QA-LoRA: Quantization-aware low-rank adaptation of large language models. In *The Twelfth International Conference on Learning Representations (ICLR)*, 2024.

Dongjie Yang, XiaoDong Han, Yan Gao, Yao Hu, Shilin Zhang, and Hai Zhao. PyramidInfer: Pyramid KV cache compression for high-throughput LLM inference. In *Findings of the Association for Computational Linguistics: ACL 2024*, Bangkok, Thailand, 2024a. Association for Computational Linguistics. URL https://aclanthology.org/2024.findings-acl.195.

June Yong Yang, Byeongwook Kim, Jeongin Bae, Beomseok Kwon, Gunho Park, Eunho Yang, Se Jung Kwon, and Dongsoo Lee. No token left behind: Reliable kv cache compression via importance-aware mixed precision quantization, 2024b. URL https://arxiv.org/abs/2402.18096.

Zhenyu Zhang, Ying Sheng, Tianyi Zhou, Tianlong Chen, Lianmin Zheng, Ruisi Cai, Zhao Song, Yuandong Tian, Christopher Ré, Clark Barrett, Zhangyang Wang, and Beidi Chen. H_2o: Heavy-hitter oracle for efficient generative inference of large language models. In *Advances in Neural Information Processing Systems 36 (NeurIPS 2023)*, 2023.

Yilong Zhao, Chien-Yu Lin, Kan Zhu, Zihao Ye, Lequn Chen, Size Zheng, Luis Ceze, Arvind Krishnamurthy, Tianqi Chen, and Baris Kasikci. Atom: Low-bit quantization for efficient and accurate llm serving. In *Proceedings of Machine Learning and Systems*, volume 6, pp. 196–209, 2024.

Fei Zuo, Zikang Zhou, Hao Cong, Xiaoyan Xi, and Ho Fai Leung. Ratequant: Optimal mixed-precision kv cache quantization via rate-distortion theory. 2026. URL https://arxiv.org/abs/2605.06675.

A Appendix: allocation histograms and uniform ladders

Per-layer bit widths (Phi-4-mini, selected arms). Gradient@3.25: 24 layers at 3 bits, 8 at 4 (layer 0 protected). Gradient@3.5: 16/16 split at 3/4 bits. KVTuner search@3.5: {2:1, 3:14, 4:17}. KVTuner search@4.0 (wide candidates): {2:3, 3:6, 4:20, 8:3}. Inverted attention@4.0: {2:2, 4:29, 8:1}.

Uniform ladder (Phi-4-mini). fp16: 2.2148; 8-bit: 2.2149; 7-bit: 2.2158; 6-bit: 2.2172; 5-bit: 2.2205; 4-bit: 2.2474; 3-bit: 2.5674; 2-bit: 7.8299 (decode CE, mean over 24 documents).

Uniform ladder (Llama-3.2-1B). fp16: 2.2565; 8-bit: 2.2566; 7-bit: 2.2574; 6-bit: 2.2623; 5-bit: 2.2814; 4-bit: 2.4303; 3-bit: 4.6093; 2-bit: 8.4793.