
Retrieval as a Quantization Multiplier: How Grounding Changes a Reader's Bit-Robustness

baa.ai Technical Report, revised 2026-10-17

I. Kennedy and T. Kennedy · baa.ai, Auckland, New Zealand

Headline. Under calibration-free uniform weight quantization (round-to-nearest, group size 64, MLX), a reader answering from retrieved context loses about half as much *relative* accuracy as the same reader answering closed-book: on Qwen3-8B, closed-book PopQA falls 12.2% at 4-bit and 22.8% at 3-bit while pooled grounded accuracy on two real corpora falls 4.9% and 12.3%. In absolute points the grounded drops are as large or larger, because grounded accuracy starts three times higher. The grounded reader's curve has the shape of the reranker's: 6-bit is certified free, 4-bit is corpus-conditional, 3-bit breaks, 2-bit destroys. On a clean single-fact grounded twin the dividend is CI-positive (+0.017 at 4-bit, +0.050 at 3-bit). A fully certified grounded-QA stack (8-bit embedder, 4-bit reranker, 6-bit reader) fits in about **8 GB** of weights; the 4B-reader stack in about **3.5 GB**.

Abstract. Quantization guidance for a language model is normally set in isolation: pick a bit-width, measure perplexity or task accuracy, ship. We show that retrieval changes the answer, and we are careful about how. Quantization's first casualty is parametric long-tail knowledge; a grounded reader receives the facts in its context, so its job is reduced to precision processing, which tolerates lower precision. On Qwen3-8B quantized to 8, 6, 4, 3 and 2 bits with calibration-free round-to-nearest group-64 quantization, closed-book PopQA falls 12.2% relative at 4-bit and 22.8% at 3-bit, while pooled grounded accuracy on two real corpora (465 judged questions) is flat at 6-bit and falls 4.9% and 12.3% relative. Because grounded accuracy starts at 0.61 against 0.20 closed-book, the absolute drops on the real corpora (−0.030, −0.075) are not smaller than closed-book (−0.024, −0.045); the matched-task test is a PopQA grounded twin, where the dividend (closed-book minus grounded degradation) is +0.017 [+0.003, +0.032] at 4-bit and +0.050 [+0.033, +0.068] at 3-bit. In clean single-fact lookup retrieval quality behaves as a switch, not a dial: with the gold fact present the reader stays at ceiling to 3-bit against 16 near-topical distractors. On real corpora, however, conditioning on whether the gold passage was retrieved does not change the bit-degradation slope, so retrieval recall@K, which we had proposed as a cheap predictor of reader bit-safety, is not one; reader bits must be certified by direct measurement. The grounded reader-size floor is 4B: a 4B reader ties a 35B, and 1.7B and 0.6B readers stay below it at 8-bit as well as at 4-bit. A reader-escalation cascade is ruled out for grounded QA: its best point is indistinguishable from the best of eleven random routings.

1. Introduction

The standard way to choose a quantization level treats the model as a closed system. You compress the weights, measure how much accuracy you lost on a benchmark the model answers from its own parameters, and set the bit-width where the loss becomes unacceptable. That procedure is correct for a model answering from memory. It is the wrong procedure for a reader inside a retrieval-augmented generation (RAG) pipeline.

Quantization does not damage a model uniformly. Its first casualty is parametric long-tail knowledge, the facts stored diffusely across the weights: compression erodes factual recall before it erodes fluency or instruction following [2, 3], and long-tail entities are exactly where parametric memory is weakest and retrieval helps most [1]. Precision loss on the computation path (following

instructions, copying a value out of context, disambiguating an entity) is far more tolerable. A grounded reader receives the facts it needs in its context window [5]. The retrieval stack has already done the remembering. That reframes the reader’s job as precision processing, which is the part of the model that survives aggressive quantization.

The consequence is that grounding is a quantization multiplier. The same reader, given the same weights at the same bit-width, tolerates more compression when it operates on retrieved context than when it answers closed-book. This report quantifies that multiplier in both relative and absolute terms, tests and rejects a cheap predictor of it, establishes the reader-size floor below which grounding stops substituting for capacity, and states one place where the mechanism does not buy what one might hope.

Relation to prior work. Yazan et al. [4] studied quantized small readers inside RAG and found that quantization costs little when the reader is grounded, which this report corroborates with a paired closed-book arm and a controlled grounded twin. Reader robustness to irrelevant or near-topical passages is documented by the RGB benchmark [7] and by Cuconasu et al. [6]; our distractor sweep extends it across bit-widths. Our cascade result is a negative instance of the FrugalGPT pattern [8] on grounded knowledge QA. The bit boundaries below are specific to calibration-free round-to-nearest quantization; calibration-based (AWQ [10], GPTQ [11]) and vector or additive quantizers (QuIP# [12], AQLM [13]) protect salient weights at 3–4 bits and may shift where the cliffs fall.

2. Setup

Reader and quantization. The core sweep uses Qwen3-8B [14] quantized to {8, 6, 4, 3, 2}-bit weights with MLX’s `mlx_lm.convert`: calibration-free affine round-to-nearest, group size 64, weight-only; activations and the KV cache stay in 16-bit. The 8-bit arm is the reference; we did not run an unquantized BF16 arm, so “lossless” below means lossless relative to 8-bit. Generation is greedy (temperature 0) with the model’s thinking mode disabled; any `<think>` span is stripped before scoring. Supporting experiments use uniform 4-bit at group size 64 for readers from 0.6B to 35B (exact checkpoints are named where they appear).

Batteries. We contrast two batteries at matched bit-depth. The *closed-book* battery is PopQA [1], a 1,000-question sample stratified by subject popularity (200 per quintile, fixed seed), chat-templated with “Answer with only the answer, no explanation”, up to 24 new tokens, scored by substring match of any gold alias. The sample contains eight duplicated question strings; Table 1 counts rows (1,000) and the twin analysis of Table 2 is keyed by question text (992). The *grounded* battery uses two internal judged question sets with reference answers: BNZ, 315 question rows (303 unique questions; seven questions are duplicated) over a 311-page bank customer-support website indexed as 1,977 chunks, in paraphrase-heavy customer voice; and 787, 162 questions over a wide-body aircraft operations manual indexed as 1,932 page parents and 9,820 380-character child windows, in verbatim technical style. Each corpus is retrieved with its certified production configuration (dense child retrieval fused with BM25 over a pool of 80, the top-3 parents returned; the 787 tier adds a cross-encoder reranker, the BNZ tier is reranker-free). Contexts were retrieved once with the embedder and reranker fixed and are byte-identical across bit-depths, so only the reader varies. For Table 1 the reader saw “Answer using ONLY the context. Reply with a short answer.” followed by the three parents and the question, up to 60 new tokens. All grounded numbers are reported on unique questions ($n = 303 + 162 = 465$); the 2026-08 draft counted the duplicated BNZ rows twice, which changed BNZ’s 8-bit accuracy from 0.597 to 0.607 and nothing else of substance.

Grounded twin and distractor stress. The *grounded twin* places the single gold sentence “*{subject}’s {relation} is {object}*” in the context of each PopQA question (recall 1, zero distractors), up to 48 new tokens, substring-scored, so closed-book and grounded accuracy are compared on the same questions and the same metric. The *distractor stress* construction takes 200 questions from six high-count relations (director, screenwriter, genre, producer, author, composer) and places the gold

fact “*The {relation} of {subject} is {object}*” plus N same-relation distractors (other subjects with their own objects) in context, shuffled with a fixed seed so the gold position is random [9], up to 24 new tokens, substring-scored. Four regimes are swept across bits: STRONG (gold, 0 distractors), MEDIUM (gold, 16), WEAK (no gold, 16), CLOSED (no context). A separate three-reader sweep over $N \in \{0, 2, 4, 8, 16, 32\}$ on 100 questions is described in §3.4.

Judge and metric. Grounded answers on BNZ and 787 are scored by a fixed judge, Qwen3-30B-A3B at 8-bit (the `mlx-community` build; a 30B mixture-of-experts model with 3B active parameters, 8-bit, same model family as the readers), prompted with the question, the reference answer and the assistant answer and asked for one word: *correct* (same key information as the reference), *partial*, or *incorrect*; answers that decline (“not found in this source”) are scored *missing*. Only *correct* counts. The judge runs greedily with a six-token budget, the prompt is identical across arms, and verdicts are cached, so the same answer always receives the same verdict. At 8-bit the verdict mix is 101 / 46 / 15 (correct / partial / incorrect) on 787 and 184 / 91 / 28 on BNZ. Because the judge shares a family with the readers and is smaller than the 35B reader of §3.5, a family-preference or ceiling bias cannot be excluded; it would compress differences between readers, not between bit-widths of the same reader. All comparisons are paired per question against the 8-bit arm with bootstrap 95% confidence intervals (2,000–5,000 resamples); SIG means the interval excludes zero, borderline means an endpoint sits at zero, ns otherwise.

Two pipelines. Tables 1–3 use the fixed minimal prompt above over precomputed contexts. The reader-size and cascade experiments (§3.5, §5) run the production pipeline end to end (production system prompt, up to 384 new tokens, same retrieval tiers, same judge). Accuracies are therefore comparable within a table but not across the two pipelines: the 4-bit Qwen3.5-4B scores 0.630 on BNZ under the production prompt while the 4-bit Qwen3-8B scores 0.568 under the minimal 60-token prompt, and that gap is a prompt and token-budget difference, not a model ranking.

3. Results

The divergence: relative and absolute

Table 1 is the central result. Closed-book PopQA degrades monotonically as bits fall. Pooled grounded accuracy is statistically free at 6-bit, borderline at 4-bit, and breaks at 3-bit. Both channels collapse at 2-bit, where generation itself fails.

Table 1: Qwen3-8B across bit-depths. Closed-book: accuracy, absolute and relative change vs 8-bit ($n=1,000$ rows; per-question hits were not persisted, so no CI). Grounded: accuracy and paired Δ vs the 8-bit arm with 95% bootstrap CI on unique questions (787 $n=162$, BNZ $n=303$, pooled $n=465$). Source: `qr_verdicts.jsonl`, recomputed 2026-10-17.

bits	closed-book PopQA	grounded 787	grounded BNZ	pooled grounded
8	0.197 (ref)	0.623 (ref)	0.607 (ref)	0.613 (ref)
6	0.198 (+0.001, +0.5%)	0.623 (± 0.000 [−.043, +.043] ns)	0.614 (+0.007 [−.020, +.033] ns)	0.617 (+0.004 [−.019, +.028] ns, +0.7%)
4	0.173 (−0.024, −12.2%)	0.611 (−0.012 [−.068, +.043] ns)	0.568 (−0.040 [−.076, −.003] SIG)	0.583 (−0.030 [−.060, .000] borderline, −4.9%)
3	0.152 (−0.045, −22.8%)	0.543 (−0.080 [−.154, −.012] SIG)	0.535 (−0.073 [−.122, −.023] SIG)	0.538 (−0.075 [−.116, −.037] SIG, −12.3%)
2	0.007 (−0.190, −96%)	0.000	0.000	0.000 (destroyed)

Read the table in both units. In relative terms, grounded reading degrades about half as much as closed-book recall: 4.9% vs 12.2% at 4-bit, 12.3% vs 22.8% at 3-bit. In absolute points it does not: the pooled grounded drops (−0.030, −0.075) are as large as or larger than the closed-book drops (−0.024, −0.045), because grounded accuracy starts at 0.61 and closed-book at 0.20. The two batteries also differ in task and metric (an LLM judge on multi-sentence answers versus substring match on single entities), so Table 1 should be read as a description of two curves on their own scales, not as a matched-task measurement of the dividend; that measurement is Table 2. The shape, however, is robust: the grounded curve is nearly identical to the reranker’s own quantization boundary measured in the same stack (4-bit free, 3-bit breaks, 2-bit destroyed), while closed-book knowledge starts dying at 4-bit and gradually. Processing components share a common cliff near

3-bit; knowledge recall degrades earlier and smoothly.

The certified operating point for a grounded reader is 6-bit, flat on both corpora (+0.004 pooled, ns), a lossless reader at 25% less memory than 8-bit (about 6.5 GB for an 8B model). 4-bit is corpus-conditional: free on verbatim 787 (−0.012 ns) but SIG-degraded on paraphrase-heavy BNZ (−0.040).

The grounding dividend on a matched task

On the PopQA grounded twin, closed-book and grounded accuracy are compared on the same 992 questions with the same substring metric (Table 2). The twin sits at ceiling (0.99) at every bit-depth while closed-book falls. Writing $\Delta_b = \text{acc}_b - \text{acc}_8$ for each channel, the dividend is $\Delta_{\text{twin}} - \Delta_{\text{closed}}$: +0.017 [+0.003, +0.032] at 4-bit and +0.050 [+0.033, +0.068] at 3-bit. This is an absolute-point quantity on a single-fact lookup task at ceiling; it establishes that the closed-book loss is not paid when the fact is in context, and it should not be read across to the real corpora of Table 1, where grounded reading is far from ceiling and does degrade.

Table 2: The grounding dividend on the grounded twin (Qwen3-8B, $n=992$ unique questions, substring metric). $\Delta_b = \text{acc}_b - \text{acc}_8$ is a signed accuracy change (negative = loss); dividend = $\Delta_{\text{twin}} - \Delta_{\text{closed}}$. Differences are computed on per-question hits before rounding, so they can differ from the difference of the rounded accuracies by 0.001. Source: `expc_reader_final.json`.

bits	closed-book	twin	Δ_{closed} [CI]	Δ_{twin} [CI]	dividend [CI]
8	0.200	0.994	ref	ref	ref
6	0.202	0.993	+0.002 [−.005, +.009]	−0.001 [−.003, .000]	−0.003 [−.011, +.004]
4	0.173	0.985	−0.026 [−.040, −.014]	−0.009 [−.016, −.004]	+0.017 [+0.003, +0.032]
3	0.149	0.994	−0.050 [−.067, −.034]	0.000 [−.003, +.003]	+0.050 [+0.033, +0.068]

Does retrieval recall predict reader bit-safety? No.

The 2026-08 draft proposed retrieval recall@K as the cheap predictor of how much dividend a corpus earns, on the evidence that 787 (recall@3 0.883) is 4-bit-free while BNZ (recall@3 0.717) is 4-bit-degraded, whereas a pairwise-cosine distractor-density score reads 0.65 for both. With two corpora that correlation is unidentifiable: 787 and BNZ differ in retrieval recall and in answer style (verbatim extraction versus paraphrase) at once. We therefore tested the mechanism within each corpus. For every question we recorded whether the gold parent was in the certified top-3 (hit@3) and recomputed the paired bit-degradation on the gold-retrieved and gold-missed subsets (Table 3).

Table 3: Qwen3-8B paired degradation vs 8-bit, conditioned on whether the gold parent was retrieved (hit@3 under the certified tier). Gold presence raises accuracy but does not change the bit slope. Source: `qr_conditional_recall.json`.

corpus	subset	n	acc 8-bit	Δ 4-bit [CI]	Δ 3-bit [CI]
787	gold in top-3	143	0.678	−0.007 [−.063, +.049]	−0.077 [−.154, .000]
787	gold missed	19	0.211	−0.053 [−.212, +.105]	−0.105 [−.316, +.105]
BNZ	gold in top-3	224	0.719	−0.040 [−.085, +.005]	−0.076 [−.134, −.018]
BNZ	gold missed	79	0.291	−0.038 [−.101, +.025]	−0.063 [−.152, +.013]
pooled	gold in top-3	367	0.703	−0.027 [−.060, +.005]	−0.076 [−.123, −.033]
pooled	gold missed	98	0.276	−0.041 [−.102, +.020]	−0.071 [−.153, .000]

Gold presence does what it should to accuracy (0.70 vs 0.28 at 8-bit) and nothing to the bit slope: the 4-bit and 3-bit degradations are the same on gold-retrieved and gold-missed questions, and the BNZ-vs-787 difference at 4-bit (−0.040 vs −0.007) persists on the gold-retrieved subset,

where recall is 1 for both. The corpus difference is therefore a property of the questions and answers (paraphrase synthesis is more bit-sensitive than verbatim extraction), not of retrieval recall. Recall@K remains the right index for certifying *retrieval*; it is not a predictor of reader bit-safety, and the self-configuration rule we proposed on that basis is withdrawn (§4.2). The distractor-density score is also not a predictor; nothing cheap we measured is.

Retrieval quality is a switch, not a dial, in clean lookup

If grounding buys compression in clean lookup, does better retrieval buy more? No. The controlled bits-by-regime sweep ($n=200$, Table 4) shows that once the gold fact is present the reader is at ceiling and stays there to 3-bit. STRONG holds at 1.000 across all bit-depths. Adding 16 near-topical distractors (MEDIUM) costs nothing: MEDIUM equals STRONG at every bit-depth, dropping only to 0.990 at 3-bit. The only bit-degradation lives in the ungrounded channel: CLOSED falls from 0.235 to 0.180 across 8 to 3 bit (paired degradation 0.055 [0.015, 0.100]), a CLOSED-vs-STRONG interaction of +0.055 [0.010, 0.100]. Because STRONG has zero variance at the ceiling, that interaction is the CLOSED degradation itself, and a ceiling can hide margin loss that has not yet flipped an answer.

Table 4: Accuracy by bit-depth $\times$ retrieval regime, $n=200$, PopQA distractor construction, substring metric, 95% bootstrap CIs. Source: `rcm_results.json`.

bits	STRONG (gold, 0 distr.)	MEDIUM (gold, 16 distr.)	WEAK (no gold, 16 distr.)	CLOSED (no context)
8	1.000	1.000	0.040 [.015, .070]	0.235 [.175, .290]
6	1.000	1.000	0.040 [.015, .070]	0.240 [.185, .300]
4	1.000	1.000	0.040 [.015, .070]	0.195 [.140, .250]
3	1.000	0.990 [.975, 1.00]	0.050 [.020, .080]	0.180 [.130, .235]

This is a statement about clean single-fact lookup, and Table 3 shows how far it carries: on real multi-sentence corpora the gold-retrieved questions still degrade at 3-bit (-0.076 SIG), so “3-bit-safe” is a property of the lookup regime, not of grounded reading in general. Within that regime the design consequence stands: improving a reranker beyond the point where it puts the gold chunk in the pool buys no extra reader bits, and a noisier distractor-heavy pool imposes no penalty either. The reader distractor-tolerance sweep reinforces it: three readers (Qwen3.5-4B at 4-bit, Qwen3-30B-A3B at 8-bit, and the interference-prone Gemma-4-26B at 4-bit; $n=100$ questions, $N \in \{0, 2, 4, 8, 16, 32\}$ same-relation distractors, gold position random) hold 0.99–1.00 accuracy at every N and never output a distractor’s object (picked-distractor rate 0 at every N). Defining reader- D^* as the largest tested N at which accuracy stays within 0.01 of $N=0$, all three readers have $D^* \geq 32$. Top- k is retrieval-limited, not reader-limited: set k to maximize the chance the gold chunk is in the pool.

The reader-size floor is 4B

Bit-robustness is not the same axis as parameter capacity. Under the production pipeline, a 4B reader ties a 35B on grounded QA: pooled 0.619 versus 0.632, $\Delta(35B - 4B) = +0.013 [-0.026, +0.050]$, statistically identical at $8\times$ smaller (2.39 GB versus 19.53 GB). Shrinking below 4B degrades grounded reading itself (Table 5): a 1.7B reader is CI-separated worse by -0.176 pooled; a 0.6B reader craters by -0.411 . Closed-book knowledge keeps dropping monotonically with size (0.6B 0.055, 1.7B 0.120, 4B 0.167, 35B 0.293 on the 992-question PopQA set), confirming the dissociation: retrieval substitutes for knowledge, not for reading capacity.

All four readers in Table 5 are at 4-bit, so parameter capacity is confounded with small-model quantization damage. To separate the two we re-ran both small readers at 8-bit and, through the same current pipeline, at 4-bit (Table 6). The pipeline is stable: the 4-bit reruns match the July arms

Table 5: Grounded accuracy by reader size, pooled BNZ+787 ($n=465$ unique questions), production pipeline, fixed 30B judge, all readers uniform 4-bit g64. Readers: Qwen3-0.6B, Qwen3-1.7B, Qwen3.5-4B (dense), Qwen3.6-35B-A3B (mixture of experts, 3B active). Source: reader_floor_final.json, deploy_reader_final.json.

reader	BNZ	787	pooled	Δ vs 4B [CI]
0.6B	0.036	0.531	0.209	-0.411 [$-0.465, -0.359$]
1.7B	0.429	0.469	0.443	-0.176 [$-0.222, -0.129$]
4B (ref)	0.630	0.599	0.619	ref
35B	0.637	0.624	0.632	$+0.013$ [$-0.026, +0.050$] ns

within -0.002 [$-0.011, +0.007$]. For the 1.7B reader bits do not matter: 8-bit scores 0.447 against 0.441 at 4-bit ($+0.007$ [$-0.039, +0.052$]), and the 8-bit 1.7B is still -0.172 [$-0.222, -0.123$] below the 4-bit 4B. For the 0.6B reader the review’s concern is borne out in part: 8-bit recovers $+0.172$ [$+0.131, +0.215$] pooled, almost all of it on paraphrase-heavy BNZ ($0.033 \rightarrow 0.281$) and none on verbatim 787 ($0.531 \rightarrow 0.562$, ns), so about half of the 0.6B’s collapse at 4-bit was quantization damage rather than capacity. Even at 8-bit, however, the 0.6B sits -0.241 [$-0.293, -0.191$] below the 4-bit 4B. The floor is therefore intrinsic: no sub-4B reader at any tested precision ties the 4B, and the sub-4B readers are also the ones for which 4-bit is not free, which removes any case for shipping them.

Table 6: Sub-4B readers at 8-bit and 4-bit through the same production pipeline (pooled BNZ+787, $n=465$ unique questions, fixed 30B judge). Δ columns are paired differences with 95% bootstrap CIs; the 4B reference is Qwen3.5-4B at 4-bit (0.619). Weight sizes are the MLX builds on disk. Source: reader_floor_8bit_final.json.

reader	bits	GB	BNZ	787	pooled	Δ 8-bit – 4-bit [CI]	Δ vs 4B@4-bit [CI]
1.7B	8	1.84	0.406	0.525	0.447	$+0.007$ [$-0.039, +0.052$] ns	-0.172 [$-0.222, -0.123$]
1.7B	4	0.98	0.426	0.469	0.441	ref	-0.176 [$-0.222, -0.129$]
0.6B	8	0.64	0.281	0.562	0.379	$+0.172$ [$+0.131, +0.215$]	-0.241 [$-0.293, -0.191$]
0.6B	4	0.35	0.033	0.531	0.207	ref	-0.411 [$-0.465, -0.359$]

4. Why it matters

Ship a certified stack, not a guessed one

Every component in a grounded-QA appliance has a certified floor on its own curve. The embedder is lossless at 8-bit (4-bit degrades, 3-bit craters); the reranker is lossless at 4-bit (3-bit breaks); the reader is certified free at 6-bit. Table 7 itemizes the two stacks. Weight sizes are measured on disk; the KV-cache figures are computed from the model configurations for a 5,000-token context (three parents plus prompt and answer), with all three models resident at once.

The 8-bit-embedder, 4-bit-reranker, 6-bit-reader stack therefore fits in about **8 GB** of weights and runs in about 8–9 GB with cache and allocator headroom; where the per-corpus certificate passes, a 4-bit reader (about 4.5 GB) brings it to about 6 GB. The smallest certified reader, Qwen3.5-4B at 4-bit, gives a resident total near **3.5 GB** (the 2026-08 draft quoted 4.4 GB, which included a gigabyte of unitemized headroom) at grounded quality statistically identical to the 35B-reader stack. The allocation rule is degenerate in the best way: spend the certified floor on each stage and stop. An 8, 16, or 32 GB grounded appliance is equivalent to the minimal certified one.

Table 7: Memory budget of the two certified stacks. Embedder and reranker are both XLM-RoBERTa-large-class models (560M parameters each); their fp16 bundle is 1.7 GB on disk and the shipped 4-bit bundle 0.87 GB, from which the 8-bit and 4-bit figures are taken. KV cache per token: Qwen3-8B, $36 \text{ layers} \times 8 \text{ KV heads} \times 128 \text{ dims} \times 2 \text{ (K, V)} \times 2 \text{ bytes} = 144 \text{ KB}$; Qwen3.5-4B, $8 \text{ full-attention layers} \times 4 \text{ KV heads} \times 256 \text{ dims} \times 2 \times 2 \text{ bytes} = 32 \text{ KB}$ plus a fixed-size state for its 24 linear-attention layers. Dense index: the 787 fp16 child vectors are 23.6 MB and BNZ’s 8.1 MB; a binary top-256 index is about one sixteenth of that.

component	precision	8B stack	4B stack
embedder (XLM-R-large class)	8-bit	0.6 GB	0.6 GB
reranker (XLM-R-large cross-encoder)	4-bit	0.3 GB	0.3 GB
reader weights	6-bit Qwen3-8B / 4-bit Qwen3.5-4B	6.5 GB	2.4 GB
KV cache at 5,000 tokens	16-bit	0.7 GB	0.2 GB
dense index (binary top-256)	1-bit	<0.01 GB	<0.01 GB
resident total		8.1 GB	3.5 GB

Reader bits must be certified, not predicted

The 2026-08 draft proposed that the near-20-question sample used to certify retrieval could also choose between a 6-bit and a 4-bit reader, driven by recall@K. Two facts rule that out. First, Table 3 shows recall does not predict the bit slope. Second, twenty questions cannot resolve the effect: one question is 5 percentage points, the exact 95% interval for 14/20 spans $[0.46, 0.88]$ and for 18/20 spans $[0.68, 0.99]$, and the -0.040 BNZ degradation that is significant at $n=303$ corresponds to 0.8 questions out of 20. Reader bit-width is still a per-corpus axis, but the certificate must come from a direct paired measurement of the candidate bit-width against the 8-bit reader on the full judged set (hundreds of questions), with 6-bit as the default whenever that measurement is unavailable. A twenty-question sample can catch gross failure (3-bit or 2-bit) and nothing finer.

The grounding dividend as a reported quantity

Closed-book degradation under reader quantization is partly (at 4-bit) to largely (at 3-bit) not paid on grounded answers to the same questions when the fact is in context (Table 2). On real corpora the relative degradation halves while the absolute degradation does not shrink (Table 1). Both curves and both differences belong on every reader-compression certificate, so a buyer sees exactly which part of the quantization tax retrieval absorbs on their corpus.

5. Limitations and honest negatives

Ruled out: the reader-escalation cascade. A tempting design routes a default small reader through a confidence gate and escalates hard queries to a big reader for “big-model quality at a fraction of cost.” It does not pay on grounded QA. On the 465 unique questions, the 4B reader ties the 35B (0.619 vs 0.632, paired +0.013 [−0.026, +0.052] ns), and the disagreements are symmetric: where they differ, the 4B is right 37 times and the 35B 43. A retrieval-confidence gate (fused dense/BM25 margin plus agreement, the signal the pipeline already computes) escalating the lowest-confidence 20% reaches 0.641, a paired +0.022 [+0.004, +0.041] over 4B-only; but that is the best of an eleven-point sweep over the escalation fraction, and shuffling the gate signal gives best-of-sweep gains of +0.016 on average (97.5th percentile +0.028, permutation $p = 0.17$), so the observed gain is indistinguishable from the best of eleven random routings. The oracle that always picks the right reader reaches 0.712, which no real gate approaches. The 2026-08 draft quoted 0.614 for 4B-only because that simulation counted the twelve duplicated BNZ rows; on unique questions it is 0.619, the Table 5 value. Ship the 4B reader. The cascade’s real home is reasoning and agentic tasks, not grounded knowledge QA.

Further honest scope:

- **One quantizer.** Every bit boundary here is for calibration-free round-to-nearest group-64 weight quantization in MLX, with 16-bit activations and no BF16 arm. Calibration-based and vector quantizers [10, 11, 12, 13] protect salient weights at 3–4 bits and may move the cliffs; the qualitative multiplier (knowledge dies first, processing later) is the claim we expect to transfer.
- **One judge, same family.** The 30B judge is a Qwen3 model grading Qwen3 readers against reference answers, counting only *correct*. Self-family bias would compress reader-vs-reader gaps (§3.5); it does not affect bit-vs-bit comparisons of one reader.
- **Two pipelines, two metrics.** Closed-book and the twin use substring match; the real corpora use the judge; Table 1 and Table 5 use different prompts and token budgets. Compare within tables.
- **4B is a floor, and not only under 4-bit.** Retrieval substitutes for knowledge, not for reading capacity. Below 4B, grounded reading degrades at 4-bit (1.7B −0.176, 0.6B −0.411 pooled) and still at 8-bit (1.7B −0.172, 0.6B −0.241); the 0.6B’s 4-bit collapse was partly quantization (8-bit recovers +0.172, almost all on BNZ). A sub-4B reader is viable only on easy verbatim corpora (on 787 the 0.6B reaches 0.53–0.56 at either precision) and never as a default.
- **Switch-not-dial is proven in clean lookup only.** The controlled result uses a single-fact grounded twin at a 1.000 ceiling. On multi-sentence real corpora grounded reading sits near 0.61, gold-retrieved questions still degrade at 3-bit, and 4-bit is corpus-conditional for reasons recall does not capture.
- **Two corpora, single family and seed.** The core sweep is Qwen3-8B, single seed, on two internal corpora. The direction replicates across the grounding-dividend twin, the distractor sweep, and the multi-reader size runs, but wider architecture and corpus coverage is future work.

6. Conclusion

Quantization guidance set in isolation asks the wrong question for a retrieval reader. Grounding changes the schedule: it removes the knowledge-deletion component of quantization damage on the questions where the fact is in context, and leaves the noise-tolerant processing path exposed, which survives to about 4-bit under round-to-nearest quantization. The measured effect is about 2× smaller relative degradation on real corpora, a CI-positive grounding dividend of +0.017 at 4-bit and +0.050 at 3-bit on a matched single-fact task, and a grounded reader whose curve has the reranker’s shape rather than the knowledge curve’s. Retrieval quality is a switch in clean lookup, but retrieval

recall does not predict how a reader's bits behave on a real corpus, so reader bit-width is a per-corpus axis that must be measured, not inferred. The practical payoff is a certified grounded-QA stack in about 8 GB, and a clear boundary on what does not work: no reader-escalation cascade, and no reader smaller than 4B.

References

- [1] A. Mallen et al. When not to trust language models: Investigating effectiveness of parametric and non-parametric memories. *ACL*, 2023. [arXiv:2212.10511](#).
- [2] S. S. S. Namburi et al. The cost of compression: Investigating the impact of compression on parametric knowledge in language models. *Findings of EMNLP*, 2023. [arXiv:2312.00960](#).
- [3] Q. Wang et al. Through a compressed lens: Investigating the impact of quantization on factual knowledge recall. *arXiv:2505.13963*, 2025.
- [4] M. Yazan et al. The impact of quantization on retrieval-augmented generation: An analysis of small LLMs. *arXiv:2406.10251*, 2024.
- [5] O. Ram et al. In-context retrieval-augmented language models. *arXiv:2302.00083*, 2023.
- [6] F. Cuconasu et al. The power of noise: Redefining retrieval for RAG systems. *SIGIR*, 2024. [arXiv:2401.14887](#).
- [7] J. Chen et al. Benchmarking large language models in retrieval-augmented generation. *AAAI*, 2024. [arXiv:2309.01431](#).
- [8] L. Chen, M. Zaharia, and J. Zou. FrugalGPT: How to use large language models while reducing cost and improving performance. *arXiv:2305.05176*, 2023.
- [9] N. F. Liu et al. Lost in the middle: How language models use long contexts. *TACL*, 2024. [arXiv:2307.03172](#).
- [10] J. Lin et al. AWQ: Activation-aware weight quantization for LLM compression and acceleration. *MLSys*, 2024. [arXiv:2306.00978](#).
- [11] E. Frantar et al. GPTQ: Accurate post-training quantization for generative pre-trained transformers. *ICLR*, 2023. [arXiv:2210.17323](#).
- [12] A. Tseng et al. QuIP#: Even better LLM quantization with Hadamard incoherence and lattice codebooks. *ICML*, 2024. [arXiv:2402.04396](#).
- [13] V. Egiazarian et al. Extreme compression of large language models via additive quantization. *ICML*, 2024. [arXiv:2401.06118](#).
- [14] A. Yang et al. Qwen3 technical report. *arXiv:2505.09388*, 2025.