FIXING THE ERRORS THAT FLIP ANSWERS: OUTPUT-WEIGHTED LOW-RANK CORRECTION FOR QUANTIZED VISION-LANGUAGE MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

A small low-rank correction added to a low-bit vision-language model (VLM) can repair only part of its quantization error. We ask which error directions such a limited correction budget should repair to preserve the answers of the full-precision (FP16) teacher. Starting from the observation that flips concentrate where the teacher's answer margin is small, so that small output changes decide them, we weight the low-rank reconstruction of the quantization residual by input-activation statistics and by the second moment of an output gradient that measures how strongly each output direction moves the answer; we call this output-weighted quantization error reconstruction OQER. With the same correction storage (rank 64, $+0.4$ bits/weight) on a 3-bit LLaVA-1.5-7B, it removes 57.5% (POPE) and 33.1% (MME) of the teacher disagreements, against 34.3% and 16.6% for activation-weighted reconstruction, and at rank 32, half the extra storage, it still exceeds activation-weighted rank 64. The gain holds on top of the MBQ and QIG quantizers, on LLaVA-1.5-13B and Qwen2.5-VL-7B. In the tested settings, retaining the structure and alignment of the output weighting improves fidelity, and several output gradients are effective.

1 INTRODUCTION

Quantizing the language backbone of a VLM to 3–4 bits changes a small fraction of its answers. Because most yes/no and multiple-choice benchmarks are scored on answers, the natural fidelity metric is the *flip rate*: the fraction of items whose answer no longer matches the FP16 teacher (Dutta et al., 2024). Flips can also be addressed by keeping some layers in FP16 or by adding a low-rank correction to the quantized weights (Yao et al., 2023; Zhang et al., 2024).

Why teacher fidelity. A quantized model is deployed as a drop-in replacement for its validated FP16 teacher, so agreement with the teacher is the quantity that decides whether behaviour is kept, and it needs no labels. Accuracy aggregates beneficial and harmful answer changes, whereas fidelity directly measures whether the teacher's behaviour is preserved: restoring the teacher also restores its errors, so accuracy can stay within its resampling noise even when many flips are removed (§6). We optimise teacher agreement and report standard metrics alongside.

A low-rank correction $M \approx \Delta W = W_T - W_Q$ can repair only part of the quantization error, so it has to choose which error directions to repair. Existing reconstructions choose them by how well the correction reproduces the weights, unweighted (Yao et al., 2023), or the layer's intermediate activations, weighted by input statistics (Zhang et al., 2024; 2025). Answers, however, flip mostly where the teacher's answer margin is small (§4), so what decides them is how far the remaining error moves the margin, not how large the error is. We therefore add an output-side weight, the second moment C_g of an output gradient (by default, the gradient of the answer margin) at each linear's output, estimated from 200 unlabelled questions. This yields a weighted truncated-SVD correction with the same storage (§5). On LLaVA-1.5-7B at 3 bits it recovers 57.5% / 33.1% of teacher disagreements on POPE / MME against 34.3% / 16.6% for the activation-weighted correction, even though its unweighted weight residual is larger. We then verify the gain with a smaller budget, on other quantizers, models and readouts, and with controls on the output weighting (§6).

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.

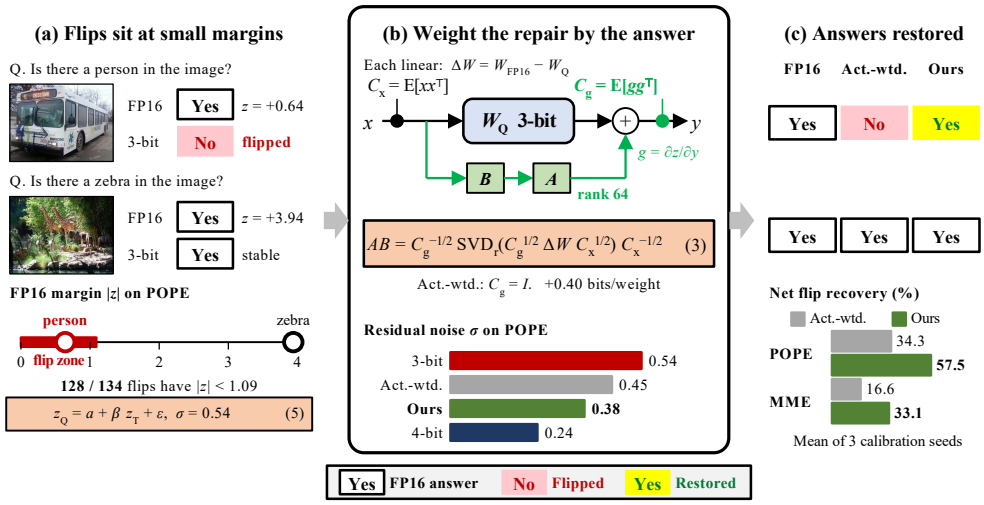

Figure 1: **Overview.** Representative items with the answer margin $\log p(\text{yes}) - \log p(\text{no})$ under FP16, 3-bit, the activation-weighted correction and OQER: a low-margin item flips at 3 bits and only OQER restores it, while a high-margin item never crosses zero. Flips concentrate where the teacher margin is small; the correction weights ΔW by C_x and by the answer-margin gradient moment C_g.

Our contributions are:

- **Output-weighted correction of the quantization residual.** We propose OQER (output-weighted quantization error reconstruction), a training-free correction that spends a limited low-rank budget on the residual of a fixed quantized checkpoint, weighting it by input-activation statistics and by output-gradient statistics that estimate how strongly each output direction moves the answer (§5).
- **More teacher answers kept at the same or a smaller correction budget.** On LLaVA-1.5-7B, it leaves fewer teacher disagreements on POPE and MME than activation-weighted reconstruction with the same correction storage, and in the tested setting a rank-32 correction, half the extra storage, still exceeds activation-weighted rank 64 (§6.1).
- **Transfer across compression settings, models and readouts.** The gain holds when the correction is added to MBQ and QIG checkpoints, on LLaVA-1.5-13B and Qwen2.5-VL-7B, and for short generated answers scored with the official parsers; we report teacher fidelity and ground-truth accuracy separately, together with where the gain is small (§6.2).
- **Controlled tests of what the output weighting needs.** A diagonal-versus-full comparison with the same gradient and an eigenvalue-preserving permutation of the output coordinates test the role of output structure and alignment, and several output gradients are shown to work, so we do not claim that the answer margin is required (§6.3.1–6.3.2).

2 Related Work

Fidelity of compressed models. Pruning damages uncommon, low-confidence examples most (Hooker et al., 2019), and flips were proposed as a compression metric for LLMs (Dutta et al., 2024). Concurrently, Wu et al. (2026) fit the same affine (shrinkage) margin model to 16 text-only LLMs, show that constants borrowed across *models* fail at 3 bits, and find no label-free repair that beats one extra bit. We differ in scope (VLM readouts, transfer across *benchmarks* at fixed model, noise and item×layer controls) and in linking the model to a repair. Saha et al. (2026) report that per-question precision selection does not beat shuffled controls in quantized VLMs, consistent with the limited, poorly generalising gains of the layer-restoration rules we tested (Appendix E).

Table 1: Design axes. Fixed W_Q: a post-hoc correction added to an already quantized checkpoint (SRR changes W_Q by quantizing after preservation; LQ-LoRA alternates quantization and low-rank fitting, then fine-tunes). Input/output: side on which the error is weighted; full: a full output moment rather than a per-channel scale (OQER: full where the output dimension allows, §5). FWSVD and LQ-LoRA weight by parameter Fisher (FWSVD row-wise, i.e. per input channel in our $y = Wx$ notation; LQ-LoRA element-wise); OQER weights by activation and output-gradient moments. †A control of this weighting on a fixed W_Q (Table 5); for FWSVD, LQ-LoRA and LASER it is inspired by, not identical to, the original weighting. ‡Run as a full method (Table 17).

Method (venue)	fixed W_Q	input	output	full	VLM
Low-rank compensation of quantization error					
LoRC (arXiv 2023)†	✓	✗	✗	✗	✗
LQER (ICML 2024)†	✓	✓	✗	✗	✗
QERA / EoRA (ICLR 2025 / 2026 W)†	✓	✓	✗	✗	✗
SRR (ICML 2026)‡	✗	✓	✗	✗	✗
Sensitivity-weighted factorisation					
FWSVD (ICLR 2022)†	✗	✓	✗	✗	✗
LQ-LoRA (ICLR 2024)†	✗	✓	✓	✗	✗
LASER (arXiv 2026)†	✗	✓	✓	✓	✓
VLM quantizers					
MBQ / QIG (CVPR 2025 / 2026)‡	✗	✓	✗	✗	✓
OQER (ours)	✓	✓	✓	✓	✓

Low-rank compensation of quantization error. Table 1 places our correction among low-rank compensation methods and VLM quantizers. Representative residual-reconstruction methods choose the correction by how well it reconstructs the weights or the intermediate activations. LoRC (Yao et al., 2023, arXiv 2023) takes an unweighted SVD of the quantization error, LQER (Zhang et al., 2024, ICML 2024) scales it by per-channel activation statistics, QERA (Zhang et al., 2025, ICLR 2025) and EoRA (Liu et al., 2026, ICLR 2026 workshop) give the optimal activation-weighted closed form, and SRR (Cho et al., 2026, ICML 2026) splits the rank between keeping a dominant subspace, quantizing the remainder, and reconstructing the rest of the error, so its quantized matrix differs from that of a fixed checkpoint. None of them weights the output directions of the residual by their effect on the answer; this is the part we add.

Sensitivity-weighted factorisation. Parameter- or loss-sensitivity has been used to guide the factorizations of weights. FWSVD (Hsu et al., 2022, ICLR 2022) weights a factorisation of the weight by row-wise parameter-Fisher importance (per input channel in our notation), LQ-LoRA (Guo et al., 2024, ICLR 2024) alternates between quantizing each weight and fitting a Fisher-weighted low-rank approximation of the residual $W - Q$, and then fine-tunes the low-rank part, and LASER (Wang et al., 2026, arXiv 2026) uses a Kronecker-factored, loss-aware weighting of the SVD with rank allocation. Our objective shares this two-sided weighted SVD, and we do not claim the closed form as new. The difference is the setting: we repair the residual of an already quantized checkpoint, without training, as an add-on to any quantizer, and we evaluate agreement with the full-precision teacher. Compared with FWSVD and LQ-LoRA, we also weight by activation and output-gradient moments rather than parameter Fisher and keep the full output moment. Applied to ΔW at fixed rank, a LASER-style objective differs from OQER only in the gradient used, and Table 5 includes such a control. In our experiments a general output gradient (a whole-prompt language-model loss) is about as effective as the answer margin (Table 7), so the answer margin is one implementation choice.

VLM quantizers and corrections. Beyond AWQ (Lin et al., 2024, MLSys 2024), GPTQ (published as OPTQ) (Frantar et al., 2023, ICLR 2023) and SmoothQuant (Xiao et al., 2023, ICML 2023), MBQ (Li et al., 2025, CVPR 2025) balances vision and language token sensitivities in calibration and QIG (Xiang et al., 2026, CVPR 2026) ranks parameters by quantization-aware integrated gradients; GRACE (Chen et al., 2026, ICML 2026) combines quantization-aware training with confidence-based distillation. Quant Experts (Jia et al., 2026, CVPR 2026) is closest to our setting: it compensates the quantization error of a VLM with a shared low-rank adapter and token-

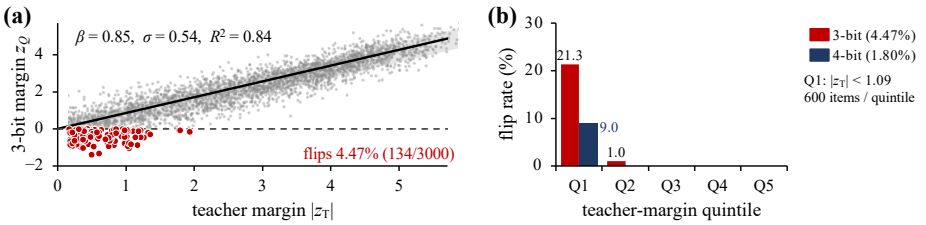

Figure 2: **Flips are sign changes of a shrunk, noisy margin** (LLaVA-1.5-7B, 3-bit, POPE-adversarial). (a) Quantized vs. FP16 margin; flips (red) concentrate where $|z_T|$ is small relative to σ. (b) Flip rate by teacher-margin quintile.

routed low-rank experts built from important channels, that is, it chooses where in the input the error is compensated. Our correction is a single static low-rank term per linear, and the weighting instead chooses which output directions its rank is spent on; the two choices are complementary. These methods decide how the weights are quantized; our correction is applied afterwards to whatever W_Q they produce, and Table 17 adds it to MBQ and QIG checkpoints.

3 Setup

Models and quantization. LLaVA-1.5-7B and LLaVA-1.5-13B (Liu et al., 2024) and Qwen2.5-VL-7B (Bai et al., 2025). Unless noted, weights of the language backbone are quantized to 3 bits with AWQ-style activation-aware scaling (Lin et al., 2024) (group size 128, calibration on 32 held-out COCO (Lin et al., 2014) images with a captioning prompt), simulated by quantize–dequantize so that FP16 and quantized models differ only in the weights (memory with real low-bit kernels: Appendix C). The vision tower and projector stay in FP16 except where stated. All experiments run on NVIDIA A100 80GB GPUs (one GPU per run) and one NVIDIA RTX A6000 48GB, with PyTorch 2.1.2 and Transformers 4.45.2 (Qwen2.5-VL: PyTorch 2.6 and Transformers 4.49); §5 gives the time needed to build the correction.

Benchmarks and readout. POPE (adversarial, random and popular splits, 3,000 yes/no items each; tables labelled POPE-adversarial use the adversarial split only) and MME (2,374 yes/no items over 14 subtasks) (Fu et al., 2023; Li et al., 2023); GQA (Hudson & Manning, 2019) and ARC (Clark et al., 2018) are used only in the margin analysis (Appendix D). We write T for the FP16 teacher, Q for the quantized model and H for a corrected model. The margin z is the log-probability of the teacher's answer minus that of the strongest other answer at the answer position; the teacher's answer is fixed once for every model, while the strongest other answer is found again in each evaluated model, so a model disagrees with the teacher exactly when its $z < 0$ (yes/no and letter readouts). Greedy decoding is used throughout; exact definitions are in Appendix B. Scoring generated answers with the official lmms-eval parsers instead of the margin gives the same ordering of the corrections on all three backbones (Appendix J).

Fidelity. A flip is an item on which Q disagrees with T. For a correction H we report the net recovery

$$R_{\text{net}} = \frac{n_{\text{recovered}} - n_{\text{new}}}{n_{\text{flips}}}, \tag{1}$$

the number of flips that H restores minus the number of new disagreements it creates, divided by the number of flips of Q. It can be negative and is not an accuracy gain; since the denominator differs between quantizers, R_{net} values of different quantizers are not a ranking of final quality (equivalent rate form in Appendix B). Differences between interventions are tested as described in Appendix A. Ground-truth accuracy is reported separately. Calibration and evaluation images are disjoint.

4 Where Answers Flip

Let z be the teacher-oriented answer margin of §3, so a model disagrees with the teacher exactly when its $z < 0$. Quantization keeps a shrunk, noisy copy of the teacher margin (Figure 2a), so

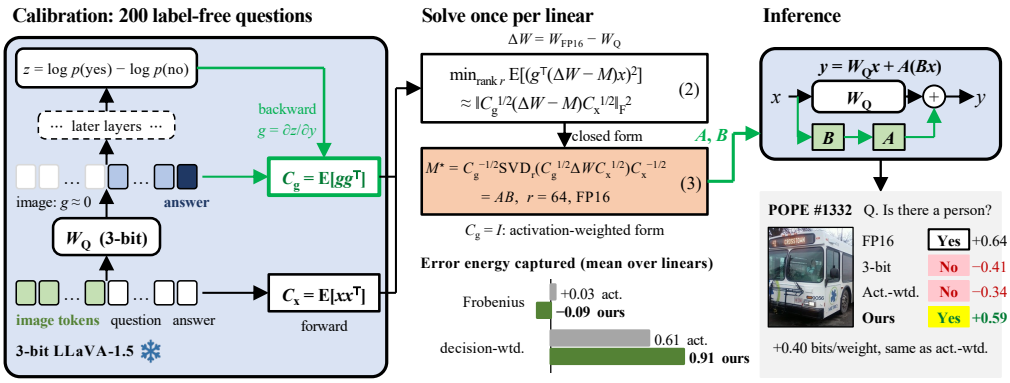

Figure 3: **The correction pipeline.** Statistics are collected once on the frozen quantized model: C_x from forward passes and C_g from one backward pass of the answer margin per calibration question. Each linear's quantization residual ΔW is then factorised by a weighted truncated SVD into FP16 factors A, B, and inference computes $y = W_Q x + A(Bx)$.

mainly items whose teacher margin is small cross zero: on POPE, the 3-bit flips fall almost entirely in the lowest teacher-margin quintile (Figure 2b). The same relation describes other benchmarks: subtasks with smaller teacher margins flip more, and a fit on one benchmark predicts the flip rate of another (Appendix D, Figure 4). The error that matters for the answers is therefore the part that moves the answer margin, which is small for most items and decisive only near the boundary.

5 Output-Weighted Low-Rank Correction

Setting. We correct each linear layer of a fixed quantized backbone by adding a rank-r FP16 term M to W_Q that targets the quantization residual $\Delta W = W_T - W_Q$ (Figure 3). Activation-weighted reconstruction chooses M to minimise the layer's local output error, weighting input directions by their second moment $C_x = \mathbb{E}[xx^\top]$. OQER additionally weights output directions by how strongly they move the answer, through the second moment $C_g = \mathbb{E}[gg^\top]$ of $g = \nabla_y z$, the gradient of the answer margin with respect to the layer output y (other output gradients are compared in §6.3.2).

What the weighting measures. After correction, $\Delta W - M$ is the weight error that remains. C_x describes how much of this error shows up on the inputs the layer actually receives, and C_g describes how much the resulting output error moves the answer scalar. To first order, the remaining error changes the answer margin of a question by

$$\delta z \;\approx\; \textstyle\sum_t g_t^\top (\Delta W - M)\, x_t,$$

summed over token positions t, so an error direction matters only if it is excited by the inputs and read by the answer.

Objective. The correction minimises

$$\min_{\operatorname{rank}(M)\le r} \left\| C_g^{1/2} \,(\Delta W - M)\, C_x^{1/2} \right\|_F^2. \tag{2}$$

This is a surrogate for the displacement of the answer margin caused by the remaining error, obtained from a first-order expansion around the quantized model. It drops cross-token and cross-layer terms and factorises the joint second moment of (x, g) as $C_x \otimes C_g$ (derivation in Appendix B). It is not the flip rate itself.

Solution and inference. For positive definite weights the exact minimiser is the weighted truncated SVD

$$M^\star = C_g^{-1/2}\, \mathrm{SVD}_r\!\left(C_g^{1/2}\, \Delta W\, C_x^{1/2}\right) C_x^{-1/2}, \tag{3}$$

where SVD_r keeps the top r singular triplets. We store it as $M = AB$ and compute $y = W_Q x + A(Bx)$ at inference. With $C_g = I$ it reduces to the activation-weighted reconstruction, and with $C_g = C_x = I$ to LoRC.

Table 2: Net recovery equation 1 at rank 64, all layers; mean over three calibration seeds (per-seed counts in Table 10). Gain: pooled over seeds, with the range of the per-seed gains.

Correction	POPE-adversarial R_{net} (%)	MME R_{net} (%)
activation-weighted	34.3	16.6
decision-weighted (OQER)	**57.5**	**33.1**
gain of OQER (per-seed range)	$+23.1^{\dagger}$ ($+16.4$ to $+26.9$)	$+16.5^{\dagger}$ ($+14.0$ to $+18.6$)
OQER at rank 32 (half correction memory)	46.0	29.6
W4 g128 (4.16 bits), reference	59.7	58.4

Implementation. C_x and C_g are collected on the frozen quantized model from 200 unlabelled calibration questions whose images are disjoint from evaluation, using forward hooks and one backward pass of z per question (activation weighting needs the forward passes only, so the calibration compute is not identical, while the correction storage is). Both moments are trace-normalised and lightly ridge-regularised before equation 3 is applied, so the implemented M solves the regularised objective exactly. For the three MLP projections we use a diagonal moment on their large side (the gate/up outputs and the down input); all other moments are full. Rank 64 on all layers adds $+0.395$ bits per quantized weight; collecting the statistics takes about 4 minutes and building the corrections about 20 minutes on one A100. Normalisation, factor shapes and the bit count are given in Appendix B, and Algorithm 1 summarises the whole procedure.

6 Experiments

Compared methods. Two kinds of comparison are made. Objective-level controls (Tables 2, 5 and 7) re-implement every weighting, including a LASER-inspired one, on the same ΔW, rank, layers, calibration questions and FP16 factors, so that only the weighting differs; iso-bit alternatives spend the same extra bits on an FP16 layer or a finer quantization grid. The full-method comparison (Table 17, Appendix L) runs MBQ and QIG with their official code and SRR with its official procedure on our quantizer. The exact definition of every compared weighting is in Appendix F.

Protocol. LLaVA-1.5-7B, W3 g128 backbone-only, rank 64 on all 32 layers ($+0.395$ bits/weight; 3.55 real bits/weight including group scales and zeros, 3.16 for plain W3), three calibration seeds of 200 COCO questions shared by all corrections; full POPE-adversarial and MME test sets.

6.1 Does Output Weighting Improve Fidelity at the Same or a Smaller Correction Budget?

Main result. At identical rank, memory and calibration questions, output-gradient weighting consistently improves on the activation-weighted correction on both benchmarks and every seed (mean net recovery 57.5% vs. 34.3% on POPE, 33.1% vs. 16.6% on MME; Table 2; pooled differences $+23.1^{\dagger}$ and $+16.5^{\dagger}$). In absolute terms, 57 instead of 88 of the 3,000 POPE-adversarial questions still disagree with the teacher, and 187 instead of 233 of the 2,374 MME questions; on MME, OQER restores 64 more disagreements than the activation-weighted correction but also breaks 18 more, and has higher point estimates than the tested near-budget alternatives (3.31–3.63 bits/weight), of which a finer group-32 grid is the strongest (Table 12). The three LLaVA seeds share one evaluation set and are calibration draws, not independent test sets.

Less weight error does not mean more preserved answers. At the same rank, the corrections order oppositely under the two criteria (seed 0; Table 22). Weight error is measured as the per-linear squared residual normalised by $\|\Delta W\|_F^2$ and averaged over linears. Among these three corrections, the unweighted SVD reduces it the most (by 6.5%) but recovers the fewest disagreements (R_{net} 15.7% / -10.0% on POPE / MME). The activation-weighted correction reduces it by 3.3% (38.1% / 15.1%). With OQER it rises by 9.1% relative to the uncorrected error, yet OQER recovers the most (54.5% / 31.9%). The gain is measured on evaluation images disjoint from the calibration questions, so it is not an artefact of the weighted objective that OQER minimises by construction.

Table 3: **Teacher fidelity across quantizers**, LLaVA-1.5-7B, W3 g128: teacher disagreement (%, lower is better) on POPE and MME. Corrections: rank 64 on all layers, mean over three calibration seeds; bits are real bits/weight. Bold: best within each quantizer. Accuracy, F1 and all rows: Table 17 (Appendix L).

		AWQ		MBQ		QIG	
correction	bits/w	POPE	MME	POPE	MME	POPE	MME
none	3.16	3.84	11.75	3.42	10.66	4.47	10.95
+ act.-weighted	3.55	2.42	9.80	2.16	9.00	2.05	10.83
+ OQER (ours)	3.55	**1.55**	**7.86**	**1.52**	**7.99**	**1.50**	**8.93**

Table 4: **Across backbones**, AWQ-style W3 g128 backbone quantization (other quantizers on LLaVA-1.5-7B: Table 3). Bits are real bits/weight; MME accuracy is per question. act.-weighted: activation-weighted reconstruction (QERA/EoRA objective). Corrections: mean over three calibration seeds. Bold: best among the quantized models of each group; gray rows are the FP16 reference, not ranked.

		real	teacher disagree. % ↓		POPE (3 splits)		MME
Backbone	correction	bits/w	POPE	MME	acc	F1	acc
	none	3.16	3.84	11.75	84.66	82.70	74.98
LLaVA-1.5-7B	+ act.-weighted	3.55	2.42	9.80	85.46	84.01	74.96
	+ OQER (ours)	3.55	**1.55**	**7.86**	**85.55**	**84.18**	**75.89**
FP16 reference	–	*16*	*0*	*0*	*85.90*	*84.59*	*78.56*
	none	3.16	2.37	10.61	85.28	83.53	**76.03**
LLaVA-1.5-13B	+ act.-weighted	3.47	2.00	8.09	85.31	83.55	75.67
	+ OQER (ours)	3.47	**1.68**	**7.09**	**85.71**	**84.09**	75.60
FP16 reference	–	*16*	*0*	*0*	*85.56*	*83.87*	*77.46*
	none	3.16	2.70	7.46	86.44	84.73	86.02
Qwen2.5-VL-7B	+ act.-weighted	3.55	2.54	6.54	86.30	84.51	86.68
	+ OQER (ours)	3.55	**1.67**	**6.15**	**87.28**	**85.87**	**87.01**
FP16 reference	–	*16*	*0*	*0*	*87.54*	*86.22*	*87.83*

Half the correction memory. At rank 32 on all layers, OQER still recovers more than the activation-weighted correction at rank 64 on both benchmarks and on every seed ($+11.7^{\dagger}$ and $+13.0^{\dagger}$ net recovery; Table 2). A rank sweep on layers 0–11 is in Table 11 (Appendix G).

Strength of the correction. Scaling either correction by $c \in \{0.25, \ldots, 1\}$ does not close the gap: a held-out dev set selects $c = 1$ for both, and at every c OQER has the higher net recovery on both benchmarks and restores more disagreements for the same number of new ones wherever the activation-weighted curve reaches (Appendix H).

6.2 Does the Correction Transfer across Quantizers, Models and Readouts?

Full-method comparison. Table 3 treats the correction as an add-on to a quantizer (accuracy and all rows in Table 17, Appendix L). On AWQ, MBQ and QIG, adding OQER removes the largest share of teacher disagreements at 3.55 bits on both benchmarks, more than the activation-weighted correction on the same quantizer, more than SRR at the same correction storage, and at least as much as a finer QIG grid that uses more bits (Table 17). All rows share model, prompt, readout and quantized modules; MBQ and QIG run with their official code, and SRR with its official procedure on our quantizer (Appendix F). Fidelity and ground-truth accuracy differ. POPE answers are binary, so a changed answer on a question the teacher gets wrong always becomes correct; the activation-weighted correction changes more of these and is 0.1–0.2 points more accurate on POPE with MBQ and QIG (not significant), whereas OQER breaks fewer of the teacher's correct answers in every split (Appendix L). MME per-question accuracy is highest with OQER on all three quantizers, significantly so against the activation-weighted correction on QIG. Unlike per-question accuracy, the official MME score (which adds a bonus for answering both questions of an image correctly) does not improve consistently with either correction (Table 21; Appendix M).

Table 5: **Weighting comparison** at equal correction storage: same W_Q, rank 64 on all layers, same calibration questions (seed 0). †: significant difference.

Correction (same W_Q)	bits	POPE-adversarial R_{net}	POPE-adversarial gain of OQER	MME R_{net}	MME gain of OQER
none (3-bit)	3.16	0.0		0.0	
LoRC (unweighted)	3.55	15.7	$+38.8^\dagger$	-10.0	$+41.9^\dagger$
LQER (diag. C_x)	3.55	3.7	$+50.7^\dagger$	16.5	$+15.4^\dagger$
Activation-weighted (QERA/EoRA objective)	3.55	38.1	$+16.4^\dagger$	15.1	$+16.8^\dagger$
diagonal C_x and C_g (FWSVD/LQ-LoRA-inspired)	3.55	28.4	$+26.1^\dagger$	3.9	$+28.0^\dagger$
diagonal C_g, full C_x (ablation)	3.55	20.1	$+34.3^\dagger$	6.1	$+25.8^\dagger$
LASER-inspired: last-24-token loss gradient	3.55	46.3	$+8.2$	27.2	$+4.7$
likelihood-gradient variant	3.55	59.7	-5.2	31.9	0.0
OQER (ours)	3.55	**54.5**		**31.9**	

On a state-of-the-art quantizer. Net recovery is not comparable across quantizers, so we test the correction within one QIG checkpoint on three seeds (POPE-adversarial and MME; full table in Appendix, Table 19). Pooled over seeds, OQER exceeds the activation-weighted correction by $+12.3^\dagger$ / $+17.3^\dagger$ net recovery, exceeds a finer group-64 grid (3.31 bits/weight) by $+28.2^\dagger$ / $+11.5^\dagger$, and is not separable from a group-32 grid that uses more bits (3.63; $+8.5$ / $+3.5$). On QIG, OQER is ahead in MME accuracy ($+1.14$pp†), while POPE-adversarial F1 does not separate the two corrections (-0.24); the gain is mainly a fidelity gain.

Other backbones. Table 4 repeats the AWQ-style comparison on LLaVA-1.5-13B and Qwen2.5-VL-7B. On both, OQER leaves the fewest teacher disagreements on POPE and MME and is ahead in POPE accuracy and F1; MME accuracy does not separate the corrections on LLaVA-1.5-13B, and on Qwen2.5-VL-7B the MME difference in net recovery is not significant.

6.3 Which Aspects of Output Weighting Matter?

6.3.1 Output structure

Table 5 holds W_Q, rank and calibration questions fixed and varies only the weighting. The unweighted SVD (LoRC), per-channel activation scaling (LQER) and the activation-weighted closed form all recover less than OQER on both benchmarks, as do diagonal output weightings. The full-versus-diagonal contrast is made with the same gradient: with the likelihood gradient, a full C_g recovers roughly three times as much as a diagonal C_g with the same full C_x on POPE and five times as much on MME, so a per-channel output scale does not reproduce the gain. With the answer-margin gradient on three seeds, the diagonal C_g also recovers significantly less than the full one on POPE for LLaVA-1.5-7B and Qwen2.5-VL-7B, and on MME for LLaVA-1.5-7B (Table 6).

6.3.2 Output gradient and alignment

Several output gradients are effective. With W_Q, C_x, rank and calibration questions fixed, we replace the answer-margin gradient by four other output gradients on three seeds (Table 7): the likelihood of the answer token, the likelihood of the teacher's own generated response (up to eight tokens), a token-averaged language-model loss over the whole prompt, and a LASER-inspired loss over the last 24 prompt tokens (the window is our choice). Every output gradient improves on activation-only weighting by a wide margin. The answer margin has the highest mean, but its lead over the whole-prompt loss, the answer likelihood and, on POPE, the response likelihood is within resampling noise. The working ingredient is therefore output-gradient weighting of the quantization error, and the answer margin is a good default rather than the only effective choice. Together with the tested diagonal ablation (§6.3.1) and the permutation control below, this supports retaining output-coordinate structure and alignment. It does not show that the output gradients are equivalent: their order depends on the setting, for example on the layers corrected and on binary versus multiple-choice readouts (Appendix I).

Table 6: Structure of the output weighting (answer-margin gradient). Net recovery (%), mean over three calibration seeds; gain of OQER pooled over seeds, †: significant. Sign control: one fixed random J for all seeds. Act. energy: captured share of the activation-weighted error $\|\Delta W C_x^{1/2}\|_F^2$. POPE: all three splits (Table 2: adversarial split only); activation-weighted and OQER rows are the runs of Table 4, so they differ from Table 2.

Model	weighting	POPE R_{net}	POPE gain of OQER	MME R_{net}	MME gain of OQER	act. energy
LLaVA-1.5-7B	activation-weighted ($C_g = I$)	37.0	$+22.7^{\dagger}$	16.6	$+16.5^{\dagger}$	0.609
	diagonal C_g	25.0	$+34.8^{\dagger}$	0.0	$+33.1^{\dagger}$	0.579
	sign control JC_gJ	25.3	$+34.4^{\dagger}$	-5.6	$+38.7^{\dagger}$	0.539
	OQER (full C_g)	**59.7**		**33.1**		0.539
Qwen2.5-VL-7B	activation-weighted ($C_g = I$)	6.0	$+32.0^{\dagger}$	12.2	$+5.3$	0.574
	diagonal C_g	11.0	$+27.0^{\dagger}$	17.7	-0.2	0.541
	sign control JC_gJ	24.1	$+13.9^{\dagger}$	15.4	$+2.1$	0.494
	OQER (full C_g)	**38.0**		17.5		0.495

Table 7: Output-gradient source for C_g (three-seed mean net recovery; same W_Q, C_x, rank 64 and questions). Gain of OQER pooled over seeds; †: significant.

C_g from	POPE-adversarial R_{net}	POPE-adversarial gain of OQER	MME R_{net}	MME gain of OQER
none (activation-weighted)	34.3	$+23.1^{\dagger}$	16.6	$+16.5^{\dagger}$
last-24-token LM loss	45.5	$+11.9^{\dagger}$	26.3	$+6.8^{\dagger}$
whole-prompt LM loss	52.5	$+5.0$	28.6	$+4.5$
teacher-response likelihood	54.0	$+3.5$	28.0	$+5.1^{\dagger}$
answer likelihood	53.7	$+3.7$	30.0	$+3.1$
answer margin (OQER)	**57.5**		**33.1**	

Output-coordinate alignment matters. Permuting the output coordinates of C_g ($C_g' = PC_gP^{\top}$, three fixed permutations) keeps its spectrum but misaligns it with the layer's outputs. Within one run, the three permutations drop net recovery to 7.5–33.6% (POPE) and 1.4–15.1% (MME), at or below the activation-weighted correction, and all six differences to OQER are significant (Table 14, Appendix I). With these permutations, keeping the eigenvalue distribution of the output-gradient moment is not enough; its alignment with the output coordinates matters. This does not show that the spectrum is irrelevant, nor that only answer-margin information matters. A permutation also moves which output channels carry weight, so it does not by itself isolate the role of the off-diagonal structure. A sign control does: JC_gJ with a random $J = \mathrm{diag}(\pm 1)$ keeps every channel's weight and the eigenvalues and changes only the signs of the off-diagonal entries. On LLaVA-1.5-7B it loses most of the gain on both benchmarks and falls below the activation-weighted correction; on Qwen2.5-VL-7B it loses about a third of the POPE gain, and MME does not resolve structure effects there (Table 6; details in Appendix K). The alignment of the off-diagonal structure therefore carries information beyond per-channel importance.

7 Conclusion

Spending a limited low-rank correction on the error directions that move the answer, estimated from output-gradient statistics, preserves substantially more of a quantized VLM's teacher answers than activation-weighted reconstruction with the same correction storage and rank, and it still does so with half the extra storage. The gain holds across calibration seeds, quantizers, three backbones and generated answers, although the corrected weights stay further from the teacher in Frobenius norm. In the tested settings, retaining the structure and alignment of the output weighting improves fidelity, and several output gradients are effective, with setting-dependent differences. Limitations, chiefly the single-token readout, the focus on fidelity rather than accuracy and the choice among output gradients, are discussed in Appendix O.

AI USE STATEMENT

We used an LLM-based coding assistant in three supporting roles. It helped us polish the wording of the manuscript, it pointed us to possibly relevant prior work, which we then read and cited only after checking it ourselves, and it helped write and run the analysis and experiment scripts following protocols we set. It also suggested some of the analyses and experimental variants we considered. Which of them to run, the acceptance criteria (recorded in advance), and the conclusions drawn from them were decided by us. Every number in the paper was recomputed by us from the stored per-item outputs, and we checked all code, citations and text produced with its help; responsibility for the paper rests with us.

REPRODUCIBILITY STATEMENT

No training or fine-tuning is involved: the correction is computed once from calibration statistics of the quantized model. §3 fixes the models, quantizers, benchmarks, margin readout, fidelity metric, hardware and library versions, and §5 states the objective and its closed-form solution. The appendices supply what is needed to rerun every experiment: the testing and resampling procedure (Appendix A), and exact margin definitions, the derivation, moment regularisation, factor shapes, bit accounting and a numerical check of the implementation (Appendix B). Each calibration seed draws 200 unlabelled COCO questions whose images are excluded from evaluation.

REFERENCES

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2.5-VL technical report. *arXiv preprint arXiv:2502.13923*, 2025.

Yanlong Chen, Amirhossein Habibian, Luca Benini, and Yawei Li. Gated relational alignment via confidence-based distillation for efficient vlms. In *International Conference on Machine Learning (ICML)*, 2026.

Yoonjun Cho, Dongjae Jeon, Soeun Kim, Moongyu Jeon, and Albert No. Preserve-then-quantize: Balancing rank budgets for quantization error reconstruction in llms. In *International Conference on Machine Learning (ICML)*, 2026.

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

Abhinav Dutta, Sanjeev Krishnan, Nipun Kwatra, and Ramachandran Ramjee. Accuracy is not all you need. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. URL https://openreview.net/forum?id=QVG7j29Sta.

Elias Frantar, Saleh Ashkboos, Torsten Hoefler, and Dan Alistarh. OPTQ: Accurate quantization for generative pre-trained transformers. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=tcbBPnfwxS.

Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Jinrui Yang, Xiawu Zheng, Ke Li, Xing Sun, Yunsheng Wu, and Rongrong Ji. MME: A comprehensive evaluation benchmark for multimodal large language models. *arXiv:2306.13394*, 2023.

Han Guo, Philip Greengard, Eric P. Xing, and Yoon Kim. Lq-lora: Low-rank plus quantized matrix decomposition for efficient language model finetuning. In *International Conference on Learning Representations (ICLR)*, 2024.

Sara Hooker, Aaron Courville, Gregory Clark, Yann Dauphin, and Andrea Frome. What do compressed deep neural networks forget? *arXiv preprint arXiv:1911.05248*, 2019.

Yen-Chang Hsu, Ting Hua, Sungen Chang, Qian Lou, Yilin Shen, and Hongxia Jin. Language model compression with weighted low-rank factorization. In *International Conference on Learning Representations (ICLR)*, 2022.

Drew A. Hudson and Christopher D. Manning. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In *CVPR*, 2019.

Chenwei Jia, Baoting Li, Xuchong Zhang, Mingzhuo Wei, Bochen Lin, and Hongbin Sun. Quant experts: Token-aware adaptive error reconstruction with mixture of experts for large vision-language models quantization. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2026.

Shiyao Li, Yingchun Hu, Xuefei Ning, Xihui Liu, Ke Hong, Xiaotao Jia, Xiuhong Li, Yaqi Yan, Pei Ran, Guohao Dai, Shengen Yan, Huazhong Yang, and Yu Wang. Mbq: Modality-balanced quantization for large vision-language models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025.

Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji-Rong Wen. Evaluating object hallucination in large Vision-Language models. In *EMNLP*, 2023.

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 *MLSys*, 2024.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft COCO: Common objects in context. In *European Conference on Computer Vision (ECCV)*, 2014.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In *CVPR*, 2024.

Shih-Yang Liu, Maksim Khadkevich, Nai Chit Fung, Charbel Sakr, Chao-Han Huck Yang, Chien-Yi Wang, Saurav Muralidharan, Hongxu Yin, Kwang-Ting Cheng, Jan Kautz, Yu-Chiang Frank Wang, Pavlo Molchanov, and Min-Hung Chen. Eora: Fine-tuning-free compensation for compressed llm with eigenspace low-rank approximation. In *ICLR 2026 Workshops (arXiv:2410.21271)*, 2026.

Sourajit Saha, Shubhashis Roy Dipta, Shaswati Saha, Nobin Sarwar, and Yuxuan Jiang. Before the flip: Measuring hidden score shifts in quantized vision language models before the answer changes for visual question answering. *arXiv preprint arXiv:2609.06922*, 2026.

Haiyu Wang, Yutong Wang, Leshu Li, Yihui Ren, and Sai Qian Zhang. Laser: Loss-aware singular-value decomposition and rank allocation for efficient low-precision vision-language models. *arXiv preprint arXiv:2606.00573*, 2026.

Zekun Wu, Swati Dhiman, and Adriano Koshiyama. Which decisions low-bit quantization breaks, and how to predict them. *arXiv preprint arXiv:2608.06564*, 2026.

Ziwei Xiang, Fanhu Zeng, Hongjian Fang, Rui-Qi Wang, Renxing Chen, Yanan Zhu, Yi Chen, Peipei Yang, and Xu-Yao Zhang. Fine-grained post-training quantization for large vision language models with quantization-aware integrated gradients. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3500–3510, 2026.

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 *ICML*, 2023.

Zhewei Yao, Xiaoxia Wu, Cheng Li, Stephen Youn, and Yuxiong He. Zeroquant-v2: Exploring post-training quantization in llms from comprehensive study to low rank compensation. *arXiv preprint arXiv:2303.08302*, 2023.

Cheng Zhang, Jianyi Cheng, George A. Constantinides, and Yiren Zhao. Lqer: Low-rank quantization error reconstruction for llms. In *International Conference on Machine Learning (ICML)*, 2024.

Cheng Zhang, Jeffrey T. H. Wong, Can Xiao, George A. Constantinides, and Yiren Zhao. Qera: an analytical framework for quantization error reconstruction. In *International Conference on Learning Representations (ICLR)*, 2025.

A Statistical Protocol

All comparisons between two interventions are paired: both are evaluated on the same items, and uncertainty is estimated by an image-cluster bootstrap (2,000 resamples of evaluation images, all questions of a resampled image kept together, 95% percentile interval). A dagger (†) marks a difference whose interval excludes zero; unmarked differences are within the range expected from resampling and are not claimed as effects. Results pooled over calibration seeds average the per-seed differences on the same resampled images, since the seeds share one evaluation set; the same image resample is applied to every method, and the three POPE splits, which share their 500 images, are resampled jointly. Net recovery is computed on the exact flip set of the quantized model, so for a single seed differences between methods are integer multiples of $1/|\mathcal{D}|$ (and of $1/(3|\mathcal{D}|)$ for three-seed means). The significance marks † refer to these intervals.

B Derivations and Implementation Details

Algorithm 1 Output-weighted low-rank correction (applied once to a quantized checkpoint; no training, no labels)

Input: teacher weights W_T and quantized weights W_Q (any quantizer), unlabelled calibration questions $\{(I_i, q_i)\}_{i=1}^n$, rank r, ridge $\lambda = 10^{-3}$. **Output:** factors (A_ℓ, B_ℓ) for every linear ℓ of the language backbone.

1 **Statistics** (on the frozen quantized model): set $C_x^\ell \leftarrow 0$, $C_g^\ell \leftarrow 0$ for every ℓ.
2 **for** $i = 1, \dots, n$ **do**
3 forward (I_i, q_i); $z_i \leftarrow \log p(\text{yes}) - \log p(\text{no})$ at the answer position
4 backward z_i; at each linear ℓ, for every token t with input x_t and output gradient $g_t = \partial z_i / \partial y_t$:
5 $C_x^\ell \mathrel{+}= x_t x_t^\top, \quad C_g^\ell \mathrel{+}= g_t g_t^\top$ (diagonal only for the large MLP sides)
6 **Correction** for every linear ℓ:
7 $\Delta W \leftarrow W_T - W_Q; \quad \widetilde{C} \leftarrow C/(\operatorname{tr} C/d) + \lambda I$ for $C \in \{C_x^\ell, C_g^\ell\}$
8 $S_g \leftarrow \widetilde{C}_g^{1/2}, \quad S_x \leftarrow \widetilde{C}_x^{1/2}; \quad U_r \Sigma_r V_r^\top \leftarrow$ top-r SVD of $S_g \, \Delta W \, S_x$
9 $A_\ell \leftarrow S_g^{-1} U_r \Sigma_r^{1/2}, \quad B_\ell \leftarrow \Sigma_r^{1/2} V_r^\top S_x^{-1}$ (stored in FP16)
10 **Inference:** $y = W_Q x + A_\ell (B_\ell x)$ at every linear ℓ.

Setting $S_g = I$ in line 8 gives the activation-weighted reconstruction; any other output gradient can replace z_i in line 3.

Margins. For a yes/no item i and model H, $u_i^{(H)} = \log p_H(\text{yes} \mid i) - \log p_H(\text{no} \mid i)$, where each side pools the case variants of the answer word by log-sum-exp under the full-vocabulary softmax at the answer position. The orientation $s_i = \operatorname{sign}(u_i^{(T)})$ is taken once from the teacher T and kept for every model, and $z_i^{(H)} = s_i u_i^{(H)}$. Then $z_i^{(T)} > 0$ (ties excluded), and, for the two-candidate readout, H disagrees with the teacher if and only if $z_i^{(H)} < 0$. Re-orienting each model by its own answer would make every margin positive and hide all flips. The equivalence concerns the yes/no readout; free-form generation can emit other tokens and is checked separately. For multiple-choice and first-token readouts the decision margin is $z_i^{(H)} = \ell_H(t_i) - \max_{c \neq t_i} \ell_H(c)$, with t_i the teacher's choice (ARC: over the choice letters; GQA: the teacher's first generated token, over the full vocabulary) and the rival re-found in H, so that any candidate overtaking t_i is a flip. A fixed teacher runner-up is used only inside the nearest-rival weighting of the ScienceQA pilot (§6.3.2), whose flips are still counted on the full argmax.

Signature fit and flip-rate prediction. (a, β) are ordinary least-squares estimates on teacher-oriented margins and σ is the residual standard deviation. A fit on raw margins would have a different intercept, since a raw bias a_0 becomes $s_i a_0$ after orientation. Under $\varepsilon \mid z_T \approx \mathcal{N}(0, \sigma^2)$ the per-item flip probability is equation 6, and the flip rate of a target benchmark with teacher margins $z_{T,1}, \dots, z_{T,N}$ is predicted by averaging it with the source-fitted parameters,

$$\widehat{f}_{\text{target}} = \frac{1}{N} \sum_{i=1}^{N} \Phi\Big(- \frac{\widehat{a} + \widehat{\beta}\, z_{T,i}}{\widehat{\sigma}} \Big);$$

this is how every predicted flip rate in §4 and Table 23 is computed. σ is a residual statistic of a sample and condition, and the prediction is probabilistic.

Net recovery. With $|\mathcal{D}| = fN$, $r_{\text{rec}} = n_{\text{rec}}/(fN)$ and $h_{\text{new}} = n_{\text{new}}/((1-f)N)$, the disagreement rate after the intervention is $f_H = f(1-r_{\text{rec}}) + (1-f)h_{\text{new}}$, hence

$$\frac{f - f_H}{f} = r_{\text{rec}} - \frac{1-f}{f} h_{\text{new}} = \frac{n_{\text{rec}} - n_{\text{new}}}{|\mathcal{D}|} = R_{\text{net}}.$$

For example, OQER on POPE seed 0 fixes 90 and breaks 17 of $|\mathcal{D}| = 134$: $R_{\text{net}} = 73/134 = 54.5\%$.

From weight error to margin displacement. For a single linear, $\mathbb{E}\|Ex\|_2^2 = \text{tr}(EC_xE^\top) = \|EC_x^{1/2}\|_F^2$ exactly, with the uncentred $C_x = \mathbb{E}[xx^\top]$. For the margin, we expand around the frozen quantized weights, on which x and g are measured: $z(W_Q + \Delta) \approx z(W_Q) + \sum_{\ell,t} g_{\ell t}^\top \Delta_\ell x_{\ell t}$. Taking $\Delta = W_{\text{FP16}} - W_Q$ for the teacher and $\Delta = M$ for the corrected model gives $\delta z = z^{(T)} - z^{(H)} \approx \sum_{\ell,t} g_{\ell t}^\top E_\ell x_{\ell t}$ with $E_\ell = \Delta W_\ell - M_\ell$. The squared displacement contains cross-token and cross-layer products. The per-linear surrogate keeps only the diagonal terms $\sum_t (g_t^\top E x_t)^2$ of each linear, which can over- or under-state the true value depending on whether token and layer effects cancel or accumulate. Moments are pooled over tokens, $C_x \propto \sum_{i,t} x_{it}x_{it}^\top$ and $C_g \propto \sum_{i,t} g_{it}g_{it}^\top$ over all positions of all calibration questions (not a per-question average). Their scale is irrelevant after the trace normalisation of equation 4.

Kronecker factorisation. $(g^\top Ex)^2 = \text{tr}(E\,xx^\top E^\top gg^\top)$ exactly, so $\mathbb{E}[(g^\top Ex)^2]$ depends on the joint moment $\mathbb{E}[(xx^\top) \otimes (gg^\top)]$. Approximating it by $C_x \otimes C_g$ gives $\text{tr}(EC_xE^\top C_g)$, and $\text{tr}(EC_xE^\top C_g) = \|C_g^{1/2} E C_x^{1/2}\|_F^2$ is again exact. This factorised weighting and the corresponding weighted SVD are standard (K-FAC-type factors; LASER (Wang et al., 2026) uses the same closed form with a different gradient).

Closed form, regularisation and memory. Before the SVD, each moment is trace-normalised and regularised,

$$\widetilde{C} = \frac{C}{\text{tr}(C)/d} + \lambda I, \qquad \lambda = 10^{-3} \tag{4}$$

(for a diagonal moment, the vector of diagonal entries divided by its mean, plus λ); the 11,008-dimensional outputs of the gate/up projections and the input of the down projection use diagonal moments. For positive definite weights, substituting $Z = \widetilde{C}_g^{1/2} \Delta W \widetilde{C}_x^{1/2}$ and $N = \widetilde{C}_g^{1/2} M \widetilde{C}_x^{1/2}$ preserves rank, so the problem is $\min_{\text{rank } N \le r} \|Z - N\|_F^2$, solved by $N = U_r \Sigma_r V_r^\top$ with minimum $\sum_{j>r} s_j^2(Z)$. The ridge in equation 4 makes $\widetilde{C}$ positive definite and also changes the objective; the implemented correction is the exact optimum of the regularised objective with the stated diagonal/full choice (diagonal C_x for the down projection, diagonal C_g for the gate and up projections, full elsewhere). The factors are $A = \widetilde{C}_g^{-1/2} U_r \Sigma_r^{1/2}$ and $B = \Sigma_r^{1/2} V_r^\top \widetilde{C}_x^{-1/2}$. For LLaVA-1.5-7B at $r = 64$ over the 7 linears of all 32 blocks they add $\sum r(d_{\text{in}} + d_{\text{out}}) = 159{,}907{,}840$ FP16 parameters to $6{,}476{,}005{,}376$ quantized weights, i.e. $16 \times 159{,}907{,}840/6{,}476{,}005{,}376 = 0.3951$ bits per quantized LLM weight. The vision tower, projector and embeddings are not in the denominator.

Simulated quantization. All quantized models in the accuracy and fidelity results are simulated: each weight is quantized and immediately dequantized to FP16, so the results reflect the effect of the weight precision, not of a particular low-bit kernel. Bits per weight are counted analytically, including group scales and zero points and the FP16 factors A, B of the correction. Memory with real low-bit kernels is measured in Appendix C.

Implementation check. Using the pipeline's own functions on random full and partly diagonal moments, we verified that (i) the weighted residual of the implemented M equals $\sum_{j>r} s_j^2$ to relative precision 10^{-8}; (ii) the logged decision-weighted energy equals $\sum_{j \le r} s_j^2 / \sum_j s_j^2$; and (iii) no alternative rank-r correction (activation-weighted, unweighted, built from another C_g, or 200 random perturbations of the factors) attains a higher energy under the same matrices.

Margin versus likelihood gradients. If the likelihood were renormalised over the two answers, $q = e^{\ell_t}/(e^{\ell_t} + e^{\ell_c}) = \mathrm{sigmoid}(z)$ and $\nabla_y \log q = (1-q)\nabla_y z$. The directions would coincide, but $C_g^{\mathrm{lik}} = \mathbb{E}[(1-q)^2 g^{\mathrm{m}} g^{\mathrm{m}\top}]$ would reweight items, so the corrections would still differ. Our likelihood variant instead differentiates the full-vocabulary $\log p_v$ of the quantized model's top-1 token, $\nabla_y \log p_v = \nabla_y \ell_v - \sum_w p_w \nabla_y \ell_w$, which is not proportional to the gradient of the raw margin, $\nabla_y u = \nabla_y \ell_{\mathrm{yes}} - \nabla_y \ell_{\mathrm{no}}$ (with $z = su$; the sign s cancels in $gg^\top$), in general. The contrast in §6.3.2 therefore compares both a direction and a weighting difference. Because the target is the model's own prediction rather than a label, we call its moment a Fisher-style factor rather than the true Fisher.

Captured energy. For a correction M and fixed weights, $\eta(M) = 1 - \|\widetilde{C}_g^{1/2}(\Delta W - M)\widetilde{C}_x^{1/2}\|_F^2 / \|\widetilde{C}_g^{1/2}\Delta W \widetilde{C}_x^{1/2}\|_F^2$, averaged over linears; with $\widetilde{C}_g = \widetilde{C}_x = I$ it is the Frobenius column. Under one common metric, equation 3 built with that metric maximises η for every linear, hence also on average. Table 22 reports each row under the $\widetilde{C}_g$ of its own run (margin or likelihood); the likelihood variant's 0.916 is measured under the likelihood $\widetilde{C}_g$ and is not comparable with OQER' 0.906. The permuted-C_g energies of §6.3.2 (0.53) are evaluated under the true margin $\widetilde{C}_g$. A Frobenius value of -0.091 means that the per-linear normalised squared residual, averaged over linears, exceeds that of the uncorrected error by 9.1%.

C Memory with Real Low-Bit Kernels

The accuracy and fidelity results use simulated quantization (Appendix B). To measure the memory of a deployed model, we replace every linear of the LLaVA-1.5-7B language backbone by a real GPTQ-format kernel of auto-gptq 0.7.1 with group size 128 and asymmetric zero points, the storage format of our quantizer (CUDA kernel for 3 bits, CUDA and exllamav2 kernels for 4 bits), and store the correction as FP16 factors A, B of rank 64 next to each quantized matrix. We record the peak allocated memory on one A100-SXM4-80GB while answering one POPE prompt (607 tokens, 576 of them image tokens) and generating 64 further tokens at batch size 1.

Table 8 shows that the 3-bit model with the correction needs 4.7 GB, about a third of the 13.6 GB of the FP16 model and less than the 4-bit models without correction. The correction itself adds 0.3 GB, matching its $+0.395$ bits/weight; the activation-weighted correction has the same rank and shapes and therefore the same memory. Computed as a separate, unfused branch $A(Bx)$, the correction also adds 3–7 ms per decoded token on every kernel, again identically for both corrections; we make no speed claim.

Table 8: Memory of LLaVA-1.5-7B with real kernels (A100-SXM4-80GB, batch 1): peak allocated memory for one 607-token POPE prompt and 64 generated tokens. Correction: rank-64 FP16 factors on every language-backbone linear.

weights (kernel)	correction	bits/w	peak memory (GB)
FP16 (cuBLAS)	–	16	13.6
W4 g128 (exllamav2)	none	4.16	4.8
	rank 64	4.55	5.1
W4 g128 (CUDA)	none	4.16	4.9
	rank 64	4.55	5.2
W3 g128 (CUDA)	none	3.16	4.5
	rank 64	3.55	4.7

D A Margin Model of Quantization Flips

Let z denote the teacher-oriented answer margin (§3). Across 3- and 4-bit quantization, POPE, MME, GQA and ARC and three VLM families, the quantized margin is well described by

$$z_Q \;=\; a + \beta\, z_T + \varepsilon, \qquad \varepsilon \approx \mathcal{N}(0, \sigma^2), \tag{5}$$

with R^2 of 0.68–0.96; the same form was reported concurrently for text-only LLMs by Wu et al. (2026). A flip is a sign change of equation 5. On POPE-adversarial with 3-bit AWQ, the fit gives slope $\beta = 0.85$ and residual standard deviation $\sigma = 0.54$ (Figure 2a). Beyond the correction experiments, this analysis also uses InternVL2.5-8B, GQA testdev (12,578 questions; margin on the first generated token, flips counted on the full answer) and ARC-Easy/Challenge (text-only, letter margin).

Fitting equation 5 per benchmark and quantizer (Table 9; ordinary least squares on teacher-oriented margins, with intercept a, slope β on the teacher margin, and residual standard deviation σ; $\beta < 1$ means the margin shrinks): Residuals are near-Gaussian (excess kurtosis 0–1.5) with mild heteroscedasticity. The slope β is 0.76–0.87 at 3 bits and ≈ 1.0 at 4 bits; the noise scale σ is 0.41–0.60 for 3-bit backbone-only quantization on yes/no readouts, 0.82 when the vision tower is also quantized, and 0.25 at 4 bits. At 2 bits (MME) $\beta = 0.09$ and 53% of answers flip: the teacher margin no longer predicts the quantized margin on this readout.

Table 9: Fit of equation 5 and the flip rate it predicts (LLaVA-1.5-7B, backbone only unless noted). Computed from the evaluation-harness margins used for the signature analysis, which differ slightly from the readout of the correction experiments (e.g. POPE W3 flips 4.63% here vs. 4.47% in §6; W4 2.07% vs. 1.80%).

Benchmark	Quantizer	β	σ	R^2	Flip obs.	Flip pred.
POPE-adv	W3 g128	0.855	0.536	0.84	4.63	4.42
POPE-adv	W3 g64	0.832	0.413	0.89	3.30	3.92
POPE-adv	W3 g128 + vision	0.822	0.823	0.68	8.40	8.14
POPE-adv	W4 g128	1.015	0.253	0.96	2.07	1.73
MME	W3 g128	0.762	0.605	0.72	11.72	13.67
MME	W2 g128	0.092	2.34	0.00	52.9	52.0
GQA	W3 g128	0.866	0.713	0.81	15.97	16.94

One signature across benchmarks. Assuming the residual is conditionally Gaussian with constant variance, the flip risk of an item follows from its teacher margin,

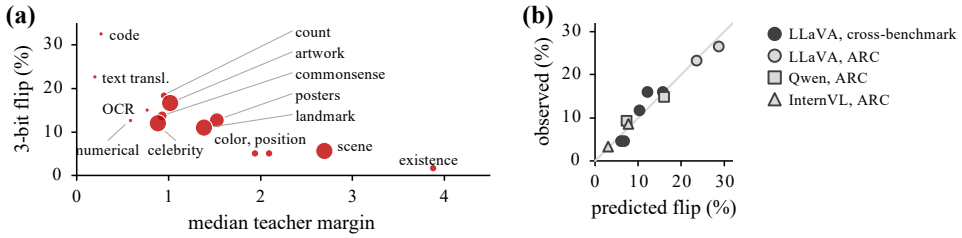

Figure 4: **One signature across benchmarks** (LLaVA-1.5-7B, 3-bit). (a) MME subtasks: median teacher margin vs. flip rate. (b) Predicted vs. observed cross-benchmark flip rate (harness margins, 139 POPE flips vs. 134 in Figure 2).

$$P(\text{flip} \mid z_T) \approx \Phi\Big(-\frac{a + \beta z_T}{\sigma}\Big). \tag{6}$$

We fit a, β, σ on a source benchmark and average these probabilities over the target benchmark's teacher margins (Appendix B). The mild heteroscedasticity limits this assumption, and a high R^2, which large margins dominate, does not by itself guarantee accurate prediction near the boundary. Applied across benchmarks, this predicts the flip rate from the target's teacher margins: POPE→MME 10.3% (observed 11.7), MME→POPE 6.1% (4.6), MME→GQA 15.8% (16.0; Figure 4b); on ARC with three families and three quantizers all 14 cross-task predictions are within 3.7pp, where transferring the flip rate directly errs by up to 15pp. The parameters' absolute scale depends on the readout, so what we observe is that the fitted parameters transfer across the benchmarks tested for a fixed model and readout, not that they are invariant in general.

Structured, not isotropic, error. Gaussian noise with the same per-linear RMS as the 3-bit error yields $\beta = 0.09$ and 42–49% flips on POPE, against 0.85 and 4.5% for quantization: consistent with quantization error concentrating in low-activation directions. Further structure checks (residual correlation across quantized models) are in Appendix E (Figure 5).

E Error Structure and Layer Restoration

This analysis explains why we correct every layer at one rank instead of selecting layers, but the method does not depend on it. The model residual ε is correlated at $\rho = 0.88$ across independently calibrated 3-bit models and at $\rho \approx 0$ between 3- and 4-bit, consistent with a residual set by the rounding grid rather than by the calibration data (Figure 5b).

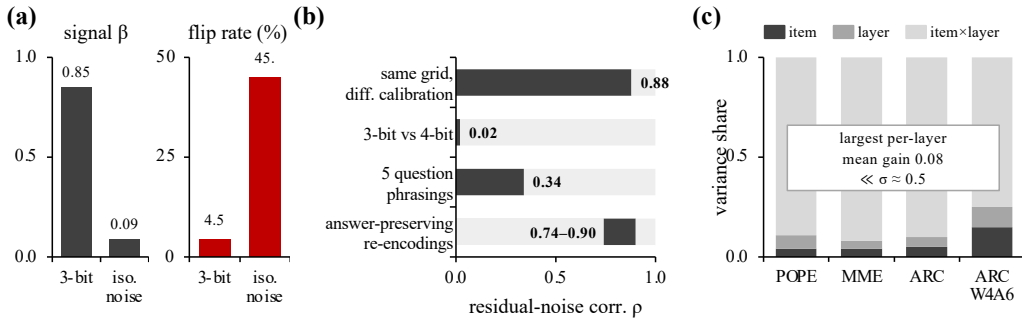

Figure 5: **Quantization error is structured.** (a) Isotropic noise of equal energy: $\beta = 0.09$, 45% flips, vs. 0.85, 4.5% for quantization. (b) Residual correlation across quantized models. (c) Item×layer restoration: interaction holds 75–92% of variance; the largest per-layer gain (0.08) is far below σ.

Keeping layers in FP16 recovers flips but re-draws part of the noise for every item, so it also creates new disagreements. Restoring the four layers with the highest single-layer restoration counts recovers about half of the flips on POPE and MME but gives $R_{\text{net}} = +23.1$ and $+24.7$ (and only $+10.6$ and $+11.3$ on images disjoint from the selection), because new disagreements offset 53–76% of the recoveries. Per-item effects add up across layers ($r \geq 0.99$) but are dominated by item×layer interaction (75–92% of variance), and the largest per-layer mean gain is 0.08 against $\sigma \approx 0.5$. With the samples and selection rules we tested, selection gains did not generalise well to held-out images. A large interaction share does not exclude structure that another selection rule could exploit, so we do not claim that layer selection cannot work.

F Compared Methods in Detail

All low-rank baselines are applied to the same quantization error ΔW with the same rank, layers, calibration questions and FP16 factors, so that only the weighting differs: LoRC (unweighted), LQER (per-channel activation scaling), the activation-weighted closed form shared by QERA and EoRA, diagonal-moment weightings inspired by FWSVD and LQ-LoRA (diagonal C_x and C_g, and diagonal C_g with full C_x; both use the likelihood gradient of §6.3.2), a LASER-inspired control that differs from OQER only in using a token-averaged loss gradient, the answer-likelihood variant (full-vocabulary log-likelihood of the quantized model's top-1 token), and OQER. Iso-bit alternatives spend the same $+0.40$ bits/weight differently: one FP16 layer (chosen by single-layer restoration count or by position), or a finer AWQ group size (64 or 32). Reference points are 3-bit and 4-bit without correction. Tables 2, 5 and 7 are objective-level controls: every weighting, including the LASER-inspired one, is re-implemented on the same ΔW, rank and calibration questions. Table 17 is a full-method comparison in which MBQ and QIG run with their official code. SRR follows its official procedure (rank split $k^\star$, Cholesky scaling of C_x, and one joint rank-64 refit of the preserved part and the quantization error, as in the official implementation) with our AWQ-style 3-bit quantizer applied to the residual weight, so that all rows share the quantizer; Table 20 also lists a split-refit variant that fits only the quantization error with rank $64 - k^\star$.

G PER-SEED RESULTS, RANK SWEEP AND ISO-BIT ALTERNATIVES

Relation to 4-bit. For calibration seed 0, at 3.55 real bits/weight (4-bit: 4.16) the corrected model is not separable from 4-bit on POPE-adversarial (disagreement 2.03% vs. 1.80%) and clearly below it on MME (-26.5pp net recovery).

Table 10 gives the per-seed counts behind Table 2.

Table 10: Per-seed counts for Table 2: net recovery equation 1 at rank 64, all layers, three calibration seeds (counts out of 134 POPE / 279 MME flips). Decision-weighted = margin gradient (OQER). rec/new: flips restored / new disagreements created; ΔGT: change in ground-truth accuracy relative to the uncorrected 3-bit model (pp).

		POPE-adversarial			MME		
Seed	Correction	rec/new	R_{net} (%)	ΔGT	rec/new	R_{net} (%)	ΔGT
0	activation-weighted	68/17	38.1	+0.23	103/61	15.1	+0.08
0	decision-weighted	90/17	54.5	+0.43	184/95	31.9	+0.97
1	activation-weighted	77/33	32.8	+0.20	105/55	17.9	−0.51
1	decision-weighted	101/21	59.7	+0.33	162/60	36.6	+1.26
2	activation-weighted	69/26	32.1	+0.23	105/58	16.8	+0.38
2	decision-weighted	104/26	58.2	+0.27	159/73	30.8	+0.51
mean	activation-weighted		34.3			16.6	
mean	decision-weighted		**57.5**			**33.1**	
gain of OQER over activation-weighted, seed 0			$+16.4^{\dagger}$			$+16.8^{\dagger}$	
W4 g128 (4.16 bits), reference		109/29	59.7	+0.33	225/62	58.4	+2.23

Table 11 varies the rank of the correction on layers 0–11, for the activation-weighted correction, OQER and the likelihood-gradient variant.

Table 11: Rank sweep on layers 0–11 (seed 0): net recovery (%) and paired gradient- minus activation-weighted difference. Last three rows: likelihood-gradient variant. r: rank of the correction; act.-w.: activation-weighted; diff: OQER minus act.-w. (pp), † significant.

	POPE-adversarial				MME			
r	LoRC	act.-w.	OQER (ours)	diff	LoRC	act.-w.	OQER (ours)	diff
8	0.0	11.9	1.5	−10.4	0.7	3.9	12.5	$+8.6^{\dagger}$
32	3.0	17.9	23.9	+6.0	8.6	14.0	25.4	$+11.5^{\dagger}$
64	6.0	17.2	38.1	$+20.9^{\dagger}$	9.0	16.8	32.3	$+15.4^{\dagger}$
8 (lik.-grad.)	0.0	11.9	3.0	−9.0	0.7	3.9	16.1	$+12.2^{\dagger}$
32 (lik.-grad.)	3.0	17.9	18.7	+0.7	8.6	14.0	22.6	$+8.6^{\dagger}$
64 (lik.-grad.)	6.0	17.2	31.3	$+14.2^{\dagger}$	9.0	16.8	18.3	+1.4

Table 12 spends the extra bits on alternatives to a correction: keeping one layer in FP16 or quantizing with a finer grid.

Table 12: Iso-bit alternatives on the same 3-bit model (seed 0): net recovery, residual scale σ, and paired gain of OQER. Bits are real bits/weight including group scales and zeros. "One FP16 layer, Lk": layer k is kept in FP16 and the others stay at 3 bits; σ: residual standard deviation of equation 5; gain of OQER: net recovery of OQER minus that of the alternative (pp), † significant.

		POPE-adversarial			MME		
Alternative	bits	R_{net}	σ	gain of OQER	R_{net}	σ	gain of OQER
one FP16 layer, L11	3.56	12.7	0.500	$+41.8^\dagger$	4.3	0.573	$+27.6^\dagger$
one FP16 layer, L10	3.56	0.7	0.533	$+53.7^\dagger$	13.3	0.554	$+18.6^\dagger$
one FP16 layer, L0	3.56	1.5	0.536	$+53.0^\dagger$	4.3	0.590	$+27.6^\dagger$
3-bit, group 32	3.63	43.3	0.377	$+11.2$	13.3	0.528	$+18.6^\dagger$
3-bit, group 64	3.31	25.4	0.413	$+29.1^\dagger$	6.8	0.593	$+25.1^\dagger$
4-bit	4.16	59.7	0.244	-5.2	58.4	0.268	$-26.5^\dagger$

H Correction Strength

A stronger correction restores more disagreements but also creates new ones, so a gain could come from strength rather than direction. We therefore scale both corrections, $W_Q + cM$ with $c \in \{0.25, 0.5, 0.75, 1\}$. The strength was chosen per method, before testing, on 3,000 held-out COCO questions: both select $c = 1$ (dev disagreement 2.23% for the activation-weighted correction and 1.63% for OQER), so the comparison at the selected strength is the one of Table 2. Table 13 gives the full curve. At every c, OQER has the higher net recovery on both benchmarks. On POPE, OQER also restores more than the activation-weighted correction at the same number of new disagreements at every c, interpolating the latter's curve. On MME the same holds for $c \leq 0.75$; at $c = 1$ OQER creates 76 new disagreements, more than the activation-weighted correction does at any strength (at most 58), so the two curves cannot be compared at that point. OQER at $c = 0.75$ restores more (144 vs. 104) and creates fewer new disagreements (50 vs. 58) than the activation-weighted correction at $c = 1$.

Table 13: Correction strength $W_Q + cM$ (AWQ-style W3, rank 64, mean over three calibration seeds). Restored: disagreements of the 3-bit model (346 POPE, 279 MME) that now match the teacher; new: teacher-agreeing answers that now differ; $R_{\text{net}} = 100\,(\text{restored} - \text{new})/|\mathcal{D}|$. $c = 1$ is the full correction, which the held-out COCO dev set selects for both methods.

		POPE			MME		
correction	c	restored	new	R_{net}	restored	new	R_{net}
activation-weighted	0.25	57.7	5.3	15.1	36.7	10.3	9.4
	0.5	111.7	15.0	27.9	60.3	22.0	13.7
	0.75	151.7	28.0	35.7	87.0	40.0	16.8
	1	185.3	57.3	37.0	104.3	58.0	16.6
	0.25	81.7	4.3	22.4	53.0	12.7	14.5
	0.5	148.0	12.0	39.3	106.7	29.7	27.6
	0.75	207.0	25.3	52.5	143.7	50.3	33.5
	1	254.3	47.7	59.7	168.3	76.0	33.1

I Choice of Output Gradient: Details

Permutation control (seed 0, registered before running): $C'_g = PC_gP^\top$ with three fixed random permutations P of the output coordinates, the same W_Q, C_x, rank 64 and statistics as OQER. Table 14 gives each arm. In this run OQER reaches 55.2% / 30.8% (POPE / MME; the separate seed-0 run of Table 10 gives 54.5% / 31.9%), and the decision-weighted energy falls from 0.91 to 0.53 under permutation.

Table 14: Permuted output weighting (seed 0). Net recovery (%); gain of OQER in the same run, †: significant. The activation-weighted row is the seed-0 run of Table 2.

weighting	POPE-adversarial R_{net}	POPE-adversarial gain of OQER	MME R_{net}	MME gain of OQER
activation-weighted ($C_g = I$)	38.1		15.1	
permuted C_g, P_1	33.6	$+21.6^{\dagger}$	11.1	$+19.7^{\dagger}$
permuted C_g, P_2	31.3	$+23.9^{\dagger}$	1.4	$+29.4^{\dagger}$
permuted C_g, P_3	7.5	$+47.8^{\dagger}$	15.1	$+15.8^{\dagger}$
OQER (ours, same run)	**55.2**		**30.8**	

The answer-likelihood variant ties with the margin on seed 0 but is significantly behind on seed 1 (Table 15) and on layers 0–11 (Table 11). On 512 ScienceQA multiple-choice questions the order reverses: a likelihood weighting (teacher's choice, softmax renormalised over the choice letters) is ahead of a nearest-rival margin weighting (teacher's choice minus the teacher's fixed runner-up; flips still counted on the full argmax), 43.2% vs. 32.6% (-10.5pp†; activation-only 24.2%). The yes/no likelihood variant differentiates the full-vocabulary log-likelihood of the quantized model's own top-1 token; its gradient is not proportional to the margin gradient, so the comparison mixes gradient direction and item weighting (Appendix B).

Table 15: Likelihood-gradient variant on the three calibration seeds of Table 2 (rank 64, all layers). rec/new: flips restored / new disagreements; ΔGT: change in ground-truth accuracy relative to the uncorrected 3-bit model (pp).

Seed	POPE-adversarial rec/new	POPE-adversarial R_{net}	POPE-adversarial ΔGT	MME rec/new	MME R_{net}	MME ΔGT
0	98/18	59.7	+0.07	162/73	31.9	+0.63
1	93/25	50.7	−0.07	161/86	26.9	+0.46
2	95/27	50.7	−0.07	151/64	31.2	+0.38
mean		53.7			30.0	

J Generated Answers Scored with the Official Parsers

The margin readout decides between the two answer tokens; Table 16 instead lets each model generate up to eight tokens and scores the text with the official parsers. The ordering is unchanged. On all three backbones, OQER has the fewest disagreements with the teacher and changes the fewest teacher-correct answers on both benchmarks. On LLaVA-1.5-7B and Qwen2.5-VL-7B it also has the highest accuracy; on LLaVA-1.5-13B the accuracies of the three quantized models lie within 0.25 points of each other.

Table 16: Generated answers (greedy, at most 8 tokens; official POPE and MME parsers), AWQ-style W3, seed 0; the teacher is the FP16 model of each backbone. Teacher-right broken: share of items the FP16 model answers correctly that the compressed model gets wrong. Bold: best among the quantized models of each backbone; gray rows are the FP16 reference, not ranked.

	real	POPE				MME		
correction	bits/w	disagree.	acc	F1	T-right broken	disagree.	acc	T-right broken
LLaVA-1.5-7B								
FP16 reference	*16*	*0*	*85.92*	*84.63*	*0*	*0*	*78.56*	*0*
none	3.16	3.96	84.59	82.61	3.08	11.92	74.98	9.81
+ activation-weighted	3.55	2.36	85.46	83.94	1.64	10.15	74.89	8.74
+ OQER (ours)	3.55	**1.67**	**85.63**	**84.22**	**1.14**	**8.30**	**75.99**	**6.86**
LLaVA-1.5-13B								
FP16 reference	*16*	*0*	*85.57*	*83.89*	*0*	*0*	*77.17*	*0*
none	3.16	2.36	85.28	83.53	1.55	11.67	**75.53**	8.41
+ activation-weighted	3.47	1.88	**85.49**	**83.79**	1.14	9.10	75.44	6.82
+ OQER (ours)	3.47	**1.71**	85.43	83.69	**1.08**	**7.75**	75.36	**6.06**
Qwen2.5-VL-7B								
FP16 reference	*16*	*0*	*87.54*	*86.22*	*0*	*0*	*87.24*	*0*
none	3.16	2.67	86.43	84.72	2.16	8.80	85.38	5.79
+ activation-weighted	3.55	2.50	86.18	84.35	2.21	7.33	86.44	4.59
+ OQER (ours)	3.55	**1.72**	**87.24**	**85.82**	**1.15**	**6.91**	**86.82**	**4.10**

K Structure Controls with the Answer-Margin Gradient

Table 6 varies only the structure of the margin-gradient moment C_g, with W_Q, C_x, rank 64 on all layers and calibration questions fixed, on three seeds. The diagonal control keeps only $\mathrm{diag}(C_g)$, that is, the weight of every output channel. The sign control uses JC_gJ with one fixed random $J = \mathrm{diag}(\pm 1)$, the same mask for all three seeds (a single sign draw, not three): it keeps the diagonal, the eigenvalues and the trace of C_g and changes only the signs of its off-diagonal entries (the gate and up projections, whose C_g is diagonal, are unchanged). On LLaVA-1.5-7B both controls lose most of the gain on both benchmarks and fall below the activation-weighted correction. On Qwen2.5-VL-7B the sign control keeps about two thirds of the POPE recovery and the diagonal control less than a third; on MME no structure effect is resolvable, and there OQER is not significantly ahead of the activation-weighted correction either. The last column shows that the activation-weighted correction removes more of the activation-weighted error than OQER on both models, while preserving fewer answers.

L Full-Method Comparison across Quantizers and POPE Accuracy

Table 17 adds the correction to three 3-bit quantizers of LLaVA-1.5-7B and compares it with the activation-weighted correction on the same quantizer, with SRR at the same correction storage and with a finer QIG grid that uses more bits. On every quantizer OQER has the lowest teacher disagreement on both benchmarks and the highest MME per-question accuracy. On POPE with MBQ and QIG, the activation-weighted correction is 0.1–0.2 points more accurate (95% image-cluster intervals include zero). POPE answers are binary, so any change on a question the teacher answers wrongly makes it correct; the activation-weighted correction changes more of these answers, while OQER changes fewer of the teacher's correct answers in every split (Table 18).

Table 17: **Full-method comparison** on LLaVA-1.5-7B, grouped by quantizer (AWQ, MLSys 2024; MBQ, CVPR 2025; QIG, CVPR 2026; SRR, ICML 2026); bits are real bits/weight. POPE accuracy by split in Table 18. MME: per-question accuracy (official scores in Table 21). act.-weighted: activation-weighted reconstruction (QERA/EoRA objective). Corrections: mean over three calibration seeds. Bold: best in group among the quantized models; gray rows (FP16, 4-bit) are references and are not ranked.

Quantizer	correction	real bits/w	teacher disagree. % ↓ POPE	MME	POPE (3 splits) acc	F1	MME acc
AWQ	none	3.16	3.84	11.75	84.66	82.70	74.98
	+ act.-weighted	3.55	2.42	9.80	85.46	84.01	74.96
	+ OQER (ours)	3.55	**1.55**	**7.86**	**85.55**	**84.18**	**75.89**
MBQ	none	3.16	3.42	10.66	84.90	83.08	75.15
	+ act.-weighted	3.55	2.16	9.00	**85.75**	**84.32**	75.62
	+ OQER (ours)	3.55	**1.52**	**7.99**	85.64	84.19	**76.02**
QIG	none	3.16	4.47	10.95	84.28	82.01	75.53
	+ act.-weighted	3.55	2.05	10.83	**85.59**	**84.21**	74.73
	+ OQER (ours)	3.55	**1.50**	**8.93**	85.42	84.02	**75.86**
SRR	preserve-then-quantize	3.55	2.30	9.56	85.67	84.41	75.15
QIG	finer grid (g32)	3.63	1.93	9.31	85.52	83.98	75.15
FP16 reference	–	*16*	*0*	*0*	*85.90*	*84.59*	*78.56*
AWQ	*W4 g128*	*4.16*	*1.48*	*4.89*	*85.53*	*84.17*	*77.21*

Table 18: POPE accuracy (%) per split and the two components of the accuracy change relative to the teacher, LLaVA-1.5-7B, mean over three calibration seeds. Broken: teacher-correct answers the corrected model changes (of 7,731). Lucky: teacher-wrong answers it changes, which become correct (of 1,269). Over the 9,000 questions, accuracy (%) = teacher accuracy $+ 100\,(\text{lucky} - \text{broken})/9{,}000$. Split columns: POPE adversarial, random and popular; bold: better of the two corrections.

Quantizer	correction	adversarial	random	popular	broken ↓	lucky
AWQ	activation-weighted	83.42	87.10	85.87	128.7	89.3
	OQER	**83.54**	**87.17**	**85.94**	**85.3**	54.0
MBQ	activation-weighted	**83.82**	**87.20**	**86.23**	103.7	90.3
	OQER	83.69	87.13	86.11	**80.0**	57.0
QIG	activation-weighted	**83.51**	**87.19**	**86.06**	106.3	78.0
	OQER	83.31	87.11	85.83	**89.0**	45.7
FP16 teacher		84.07	87.33	86.30	0	0

OQER changes fewer of the teacher's correct answers on every quantizer, but it also reproduces more of the teacher's errors, and the second effect can outweigh the first. Over the nine runs, 947 answers agree with the teacher only under OQER and 391 only under the activation-weighted correction. The teacher is correct on 55.9% of the former and 70.6% of the latter, so on the items where the two corrections disagree, following the teacher yields little accuracy. These accuracy differences are within resampling noise and change sign across seeds on MBQ; we therefore report them but do not optimise for them.

Table 19 gives the paired gains of OQER within one QIG checkpoint, and Table 20 lists every full-method row with the official MME scores.

Table 19: Gains of OQER on one QIG checkpoint over the alternatives of Table 17, pooled over three calibration seeds; POPE columns use the adversarial split only (Table 17 aggregates three splits), so the two tables differ. R_{net}: additional share of the QIG disagreements removed, in points. †: significant.

		POPE-adversarial		MME	
OQER compared with	bits/w	R_{net}	F1	R_{net}	acc
+ activation-weighted (QERA/EoRA)	3.55	$+12.3^{\dagger}$	-0.24	$+17.3^{\dagger}$	$+1.14^{\dagger}$
QIG alone	3.16	–	$+1.27^{\dagger}$	–	$+0.34$
finer grid, group 64	3.31	$+28.2^{\dagger}$	$-0.70^{\dagger}$	$+11.5^{\dagger}$	$+0.17$
finer grid, group 32	3.63	$+8.5$	-0.01	$+3.5$	$+0.72$

Table 20: Full-method comparison, all rows (LLaVA-1.5-7B, seed 0 unless noted; *mean over three seeds). POPE-adv.: adversarial split; perc./cog./total: official MME perception, cognition and total scores; disagree. %: share of questions answered differently from the FP16 teacher.

		real	POPE-adv.		MME			disagree. %	
Method (venue)	quantizer / correction	bits/w	acc	F1	perc.	cog.	total	POPE	MME
FP16	–	16	84.07	82.92	1454.2	328.6	1782.8	0.00	0.00
AWQ (MLSys 2024)	W3 g128	3.16	83.20	81.36	1367.0	348.9	1715.9	4.47	11.75
AWQ (MLSys 2024)	W3 g64	3.31	83.60	82.53	1369.9	333.2	1703.1	3.33	10.95
MBQ (CVPR 2025)	W3 g128	3.16	83.07	81.39	1388.7	287.5	1676.2	4.07	10.66
QIG (CVPR 2026)	W3 g128	3.16	83.03	80.83	1391.4	321.1	1712.5	5.23	10.95
QIG (CVPR 2026)	W3 g32	3.63	83.53	82.11	1393.2	285.7	1678.9	2.27	9.31
QIG (CVPR 2026)	W3 g64	3.31	83.63	82.80	1397.4	285.0	1682.4	3.30	10.19
QERA / EoRA objective (ICLR 2025 / ICLR 2026 W)	AWQ W3 + activation-weighted r=64	3.55	83.43	82.09	1351.6	339.6	1691.2	2.77	9.98
SRR (ICML 2026)	preserve-then-quantize, joint refit, rank 64	3.55	83.30	82.25	1393.3	295.4	1688.6	2.83	9.56
SRR (ICML 2026), split refit	preserve-then-quantize, rank $64 - k^{\star}$ error fit	3.55	83.40	82.56	1345.2	315.0	1660.2	3.33	10.91
MBQ + activation-weighted	MBQ W3 + QERA/EoRA objective r=64	3.55	83.90	82.63	1378.1	298.2	1676.3	2.57	8.76
MBQ + OQER (ours)	MBQ W3 + decision-weighted r=64	3.55	83.67	82.30	1381.3	305.7	1687.0	1.93	7.75
QIG + activation-weighted	QIG W3 + QERA/EoRA objective r=64	3.55	83.63	82.46	1345.3	302.9	1648.2	2.43	10.99
QIG + OQER (ours)	QIG W3 + decision-weighted r=64	3.55	83.40	82.19	1377.8	323.9	1701.7	**1.73**	8.68
QIG + activation-weighted, r=128	QIG W3 + QERA/EoRA objective r=128	3.95	83.83	82.87	1373.4	328.6	1701.9	2.57	10.24
QIG + OQER (ours), r=128	QIG W3 + decision-weighted r=128	3.95	83.43	82.24	1398.3	330.0	1728.3	**1.63**	7.75
OQER (ours)*	AWQ W3 + decision-weighted r=64	3.55	83.54	82.34	1370.9	325.0	1695.9	1.90	7.86
AWQ + activation-weighted, r=128	AWQ W3 + QERA/EoRA objective r=128	3.95	83.67	82.57	1363.2	349.3	1712.4	2.73	9.27
OQER, r=128	AWQ W3 + decision-weighted r=128	3.95	83.47	82.22	1401.4	328.6	1730.0	1.73	7.88
AWQ (MLSys 2024)	W4 g128, reference	4.16	83.53	82.36	1443.1	316.4	1759.5	1.80	4.89

M Official MME Scores and Why They Can Fall While Accuracy Rises

The official score adds, over 14 subtasks, 100×accuracy and 100×accuracy+ (both questions of an image answered correctly), so every subtask carries the same weight whatever its size. The eight subtasks with at most 60 questions hold 17% of the questions but 57% of the attainable points: one question there is worth 1.7–2.5 accuracy points plus a share of the pair bonus, against 0.25–0.7 on the six large subtasks (scene, landmark, artwork, celebrity, posters, commonsense reasoning).

Table 21: Official MME scores (perception: 10 subtasks, max 2000; cognition: 4 subtasks, max 800) for the rows of Tables 17 and 4; corrections are three-seed means. Bold: best within each group among the quantized models; gray rows (FP16, 4-bit) are references, not ranked. Why these can move against per-question accuracy: Appendix M.

Quantizer (venue)	correction	real bits/w	MME official score perception	cognition	total
	none	3.16	1367.0	**348.9**	**1715.9**
AWQ (MLSys 2024)	+ activation-weighted (QERA/EoRA)	3.55	1344.2	339.8	1684.0
	+ OQER (ours)	3.55	**1370.9**	325.0	1695.9
	none	3.16	**1388.7**	287.5	1676.2
MBQ (CVPR 2025)	+ activation-weighted	3.55	1381.7	295.8	1677.5
	+ OQER (ours)	3.55	1366.1	**320.5**	**1686.6**
	none	3.16	**1391.4**	**321.1**	**1712.5**
QIG (CVPR 2026)	+ activation-weighted	3.55	1349.5	313.8	1663.4
	+ OQER (ours)	3.55	1377.0	319.5	1696.5
SRR (ICML 2026)	preserve-then-quantize, rank 64	3.55	1393.3	295.4	1688.6
QIG (CVPR 2026)	finer grid, group 32	3.63	1393.2	285.7	1678.9
FP16	–	*16*	*1454.2*	*328.6*	*1782.8*
AWQ (MLSys 2024)	*W4 g128, reference*	*4.16*	*1443.1*	*316.4*	*1759.5*
Qwen2.5-VL-7B-Instruct, AWQ-style W3 g128					
	none	3.16	1658.2	**579.6**	2237.8
Qwen2.5-VL-7B	+ activation-weighted (QERA/EoRA)	3.55	**1689.1**	572.5	2261.6
	+ OQER (ours)	3.55	1687.7	579.3	**2267.0**
FP16 (Qwen2.5-VL-7B)	–	*16*	*1690.9*	*619.6*	*2310.5*
	none	3.16	**1396.6**	269.3	**1665.8**
LLaVA-1.5-13B	+ activation-weighted (QERA/EoRA)	3.47	1380.5	**269.8**	1650.2
	+ OQER (ours)	3.47	1380.4	266.9	1647.3
FP16 (LLaVA-1.5-13B)	–	*16*	*1409.1*	*263.2*	*1672.3*

The damage and the repair sit where the score gives little weight. On LLaVA-1.5-7B, FP16 answers 72–84 more questions correctly than the 3-bit models on the large subtasks, but only 0–4 more on the small ones, so the small subtasks offer no headroom. Relative to no correction, OQER (three-seed mean) gains 25 / 12 / 23 questions on the large subtasks for AWQ / QIG / MBQ, worth $+18$ / $+6$ / $+23$ points, and changes the small subtasks by a net 2–4 questions, which already costs 38 / 22 / 13 points. The official total therefore moves by -20 / -16 / $+10$ points, while per-question accuracy rises by 0.91 / 0.34 / 0.87pp. Resampling images within subtasks gives 95% intervals of roughly ± 60 points for these score differences, all of which contain zero.

Restoring the teacher also restores its errors. MME answers are binary, so every disagreement with FP16 changes correctness, and 34–36% of the 3-bit disagreements turn a wrong teacher answer into a right one. A correction that moves the model back towards FP16 undoes some of these: for AWQ, OQER repairs 107 teacher-correct answers and breaks 50, but also undoes 62 of the lucky answers and creates 26 new ones. On cognition, where 3-bit AWQ is right on 22 questions the teacher gets wrong and wrong on 24 it gets right, the 3-bit score (348.9) exceeds FP16 (328.6); OQER brings cognition to 319–325 on all three quantizers, close to FP16, from above for AWQ and from below for MBQ.

For these reasons the main text reports per-question accuracy, which weights every question equally (Tables 4 and 17), and gives the official score here for completeness.

N Where Flips Happen: Examples on Qwen2.5-VL-7B

Figure 6 relates the flips of Qwen2.5-VL-7B to the teacher margin. Panel (a) bins all 9,000 POPE questions by $|z_T|$: disagreement with the teacher is confined to small margins (none above $|z_T| = 4$), and it is in these bins that OQER removes the most disagreements (below $|z_T| = 1$: 33% and 11%

against 41% and 19% for the activation-weighted correction). Panels (b)–(e) take one image each and contrast a question with a large teacher margin, which every model answers like the teacher, with a question on the same image whose teacher margin is small, which the 3-bit model and the activation-weighted correction flip and OQER keeps on all three calibration seeds. Such questions are few: 10 distinct POPE questions and 5 MME questions show this pattern on all seeds, and the opposite pattern (kept by the activation-weighted correction on every seed and by OQER on none) occurs for one POPE question.

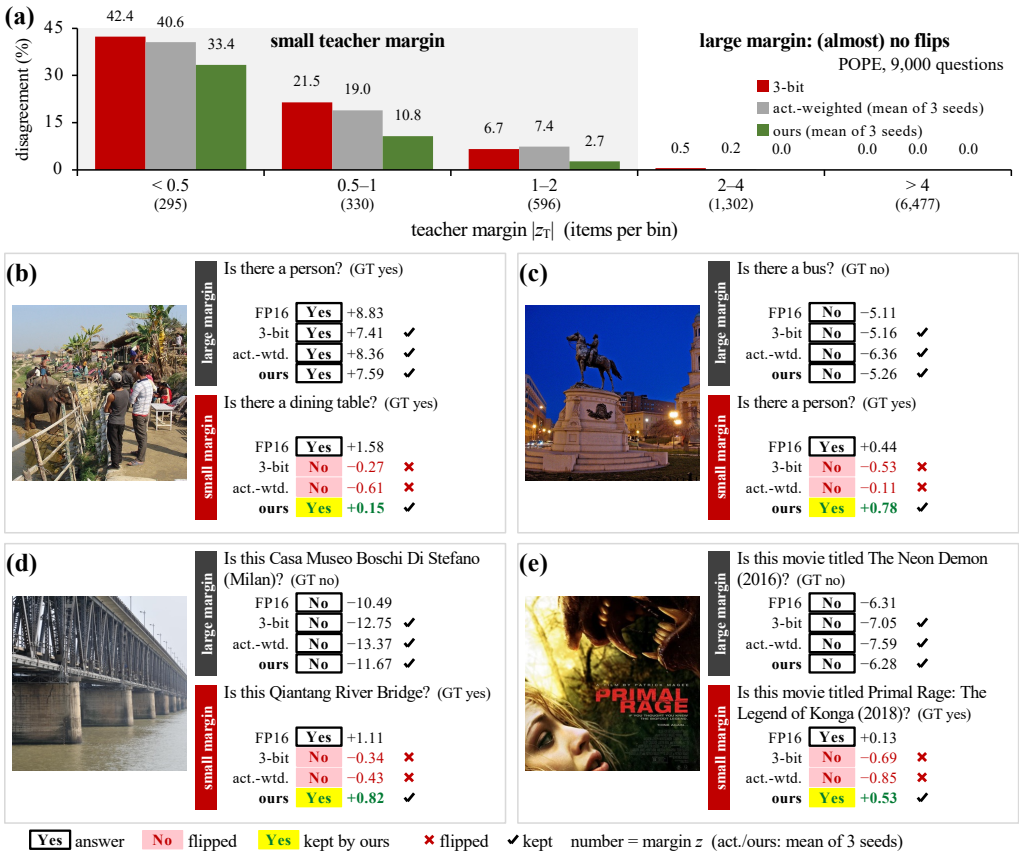

Figure 6: **Flips happen at small teacher margins** (Qwen2.5-VL-7B, AWQ-style W3 g128, rank 64). (a) Disagreement with the FP16 teacher by teacher-margin bin, POPE, three splits (bin sizes in parentheses; corrections: mean over three calibration seeds). (b)–(e) One image per panel with two questions: a large-margin question (all models keep the teacher's answer) and a small-margin question on the same image (the 3-bit model and the activation-weighted correction flip; OQER keeps the answer). Strips show the answer margin $\log p(\text{yes}) - \log p(\text{no})$ (positive = "yes"); dots are calibration seeds, bars their mean; values beyond the axis are printed. (b,c) POPE; (d,e) MME.

O LIMITATIONS

The decision direction is defined for single-token readouts; the fidelity improvement also holds under short greedy generation with the official answer parsers (Table 16), but longer free-form answers are untested, and for GQA the margin is taken on the first generated token while flips are counted on the full answer, so margin sign and answer disagreement coincide exactly only for the yes/no and letter readouts. Standard accuracy metrics barely separate the corrections (on MBQ and QIG the activation-weighted correction is marginally ahead in POPE accuracy); fidelity, not accuracy, is what we optimise. Several output gradients are effective, but their relative performance varies across settings; we do not establish a general rule for choosing among them.

P CAPTURED ERROR ENERGY

Table 22: Fraction of $\|\Delta W\|^2$ captured under the Frobenius and gradient-weighted norms (rank 64, seed 0), and net recovery. The table shows which objective each correction reduces; the decision-weighted energy does not order all arms by net recovery (e.g. diagonal C_g). Both columns are $1 -$ (squared residual)/(squared ΔW), averaged over linears; negative Frobenius values mean the squared residual grew. The decision-weighted column uses the regularised $\widetilde{C}_g$ of the run in which the row was built: margin gradient (m) or likelihood gradient (l). Values are comparable only within a column group of the same C_g. Under the margin $\widetilde{C}_g$, no rank-64 correction can exceed OQER (0.906), because equation 3 is the exact optimum for that metric.

Correction	Frobenius	decision-weighted (C_g)	POPE-adv. R_{net}	MME R_{net}
LoRC (rank 64, layers 0–11)	0.072	0.111 (m)	6.0	9.0
LoRC	0.065	0.094 (l)	15.7	−10.0
LQER (diagonal activation scaling)	0.043	0.154 (l)	3.7	16.5
activation-weighted	0.033	0.606 (m)	38.1	15.1
diagonal C_x and C_g	0.024	0.368 (l)	28.4	3.9
diagonal C_g, full C_x	−0.020	0.688 (l)	20.1	6.1
margin-gradient weighted (OQER)	−0.091	0.906 (m)	54.5	31.9
likelihood-gradient variant	−0.102	0.916 (l)	59.7	31.9

Q RESIDUAL DECOMPOSITION OF THE CORRECTED MODELS

Residual analysis. Refitting equation 5 on the corrected models (Table 23), OQER has the smallest residual on the lowest teacher-margin quintile and the largest slope β. Residual variance alone does not rank fidelity, however: on POPE, the diagonal-C_g correction has lower residual variance than activation weighting, both globally and on the lowest quintile, but more answer disagreements, and the model's predicted flip rates order these two arms the wrong way. The intercept and slope matter as well, so this analysis describes the corrected models rather than explaining the ranking of every arm.

Table 23: Fit of equation 5 per arm (seed 0). a, β, σ: intercept, slope and residual standard deviation of equation 5 fitted to the corrected model; σ_{Q1}: residual standard deviation on the lowest teacher-margin quintile. pred. %: flip rate predicted by averaging equation 6 over the benchmark; obs. %: observed flip rate, with the rate on the lowest quintile in parentheses. σ_{Q1} with β generally tracks the flip rate (exception: the likelihood variant has a slightly larger σ_{Q1} than OQER but fewer POPE flips); the global σ does not.

	POPE-adversarial						MME					
Arm	a	β	σ	σ_{Q1}	pred. %	obs. % (Q1)	a	β	σ	σ_{Q1}	pred. %	obs. % (Q1)
3-bit	0.01	0.853	0.536	0.612	4.3	4.5 (21.3)	−0.05	0.762	0.606	0.530	13.7	11.8 (36.8)
LoRC	−0.01	0.864	0.453	0.518	3.8	3.8 (18.3)	−0.02	0.759	0.661	0.572	13.9	12.9 (40.4)
LQER	0.01	0.878	0.518	0.597	4.0	4.3 (20.7)	−0.02	0.802	0.561	0.473	11.7	9.8 (34.3)
QERA / EoRA	−0.02	0.894	0.453	0.445	3.8	2.8 (13.8)	−0.05	0.802	0.532	0.468	11.9	10.0 (33.5)
diag. C_x and C_g	0.01	0.849	0.436	0.467	3.4	3.2 (15.8)	0.00	0.756	0.590	0.511	12.4	11.3 (35.8)
diag. C_g, full C_x	−0.05	0.876	0.381	0.427	3.5	3.6 (17.3)	−0.01	0.782	0.593	0.535	12.3	11.0 (35.8)
likelihood variant	0.01	0.915	0.383	0.359	2.8	1.8 (9.0)	0.00	0.818	0.483	0.430	9.7	8.0 (29.3)
OQER	0.02	0.921	0.376	0.354	2.7	2.0 (10.2)	−0.03	0.854	0.478	0.426	9.9	8.0 (28.0)
3-bit, group 64	−0.04	0.831	0.413	0.440	3.9	3.3 (16.5)	−0.05	0.811	0.593	0.497	12.8	11.0 (38.5)
3-bit, group 32	0.06	0.874	0.377	0.411	2.5	2.5 (12.7)	−0.02	0.863	0.528	0.454	10.4	10.2 (36.2)
4-bit	−0.04	1.007	0.244	0.256	2.0	1.8 (9.0)	−0.06	0.950	0.268	0.240	6.3	4.9 (23.0)