QUANTIZE ONCE, ADAPT IN PLACE: ZEROTH-ORDER OPTIMIZATION WITH DETECTABLE SIGNALS

Anonymous authors
Paper under double-blind review

ABSTRACT

Quantization reduces the memory required to store and deploy large language models, while zeroth-order (ZO) optimization avoids storing activations for backpropagation. Combining these two advantages still leaves a key question: which parameters should be optimized? We therefore study a simple constraint: keep the packed weights and quantization metadata fixed, and fine-tune only the floating-point model tensors retained in the checkpoint. We collect these tensors into the Quantization-Preserved Subspace (QPS) and propose Quantization-Preserved Zeroth-order Optimization (QPZO), which applies two-point ZO updates only in QPS. This enables adaptation without adding trainable modules or dequantizing the low-bit backbone for optimization. We support QPZO with theoretical analysis and empirical evidence. Specifically, (1) we show that an updateable output head retained in QPS yields a nonzero projected gradient whenever its aggregate gradient is nonzero, establishing the presence of task-relevant signal; and (2) because ZO accesses this signal only through finite forward-loss differences, we introduce Signal-to-Error Ratio (SER) to compare directional task-signal energy with measurement-error energy. Under the stated error bound, we prove that SER above one is the sharp threshold for guaranteeing positive expected alignment, characterizing when finite forward queries robustly preserve a useful task direction. Our theoretical analysis and extensive experiments demonstrate that QPZO offers a simple approach that substantially reduces GPU memory usage while maintaining highly competitive performance and stability.

1 INTRODUCTION

As large language models (LLMs) continue to grow in scale, the memory required to adapt them has become an increasingly important bottleneck. Weight quantization reduces this cost by storing the major model weights at low precision (Frantar et al., 2023; Lin et al., 2024), while zeroth-order (ZO) optimization removes backpropagation and estimates updates using only forward-loss evaluations (Spall, 1992; Malladi et al., 2023). These two techniques reduce memory from different sides and are naturally complementary. Combining them, however, leaves one fundamental question: which parameters should be optimized? Standard full-parameter ZO relies on continuous perturbations, applying it to an already quantized checkpoint requires restoring continuous weights for optimization, giving back much of the memory saving provided by quantization. Recent methods such as QZO (SHANG et al., 2026) and QuZO (Zhou et al., 2025) enable ZO adaptation of quantized models by updating quantization-related or low-precision parameters. Both demonstrate that ZO can operate with quantized models, yet they modify either the quantization parameters or the low-bit weights during adaptation. This leaves a simpler possibility unexplored: can an already quantized checkpoint be adapted while its quantized state remains unchanged? We therefore study a simple constraint: keep the packed low-bit weights and quantization metadata fixed, and fine-tune only the floating-point model tensors retained in the checkpoint.

These retained tensors already belong to the original network. In the quantized checkpoints we study, embeddings, normalization weights, and architecture-dependent output projections remain in floating point alongside the low-bit backbone, enabling continuous, task-driven updates through the original forward computation without adding trainable parameters. The output projection gives the most direct example: when it is retained and updateable, the standard cross-entropy objective supplies it with a task-dependent gradient through the prediction error and hidden states. Having task signal,

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.

however, does not make that signal usable by ZO. A two-point ZO optimizer never observes this gradient directly, it only observes a small loss difference between two forward queries. The remaining question is therefore whether the task signal contained in the retained parameters can still be reliably exposed through finite forward queries.

To answer this question, we introduce the **Signal-to-Error Ratio (SER)** to characterize whether task-relevant directional signals remain detectable through finite forward queries. Under the stated mean-square error bound, we prove that SER above one is the sharp threshold for guaranteeing positive expected alignment between the ZO estimate and the true gradient. With task signal present in the retained tensors and a criterion for its detectability, we call this checkpoint-native update space the **Quantization-Preserved Subspace (QPS)** and propose **Quantization-Preserved Zeroth-order Optimization (QPZO)**. QPZO is simple and consists of two steps. It first identifies QPS directly from the quantized checkpoint, and then applies standard two-point ZO updates only within QPS while keeping the low-bit weights and quantization metadata fixed. Figure 1 summarizes the overall QPS–QPZO framework and the role of SER.

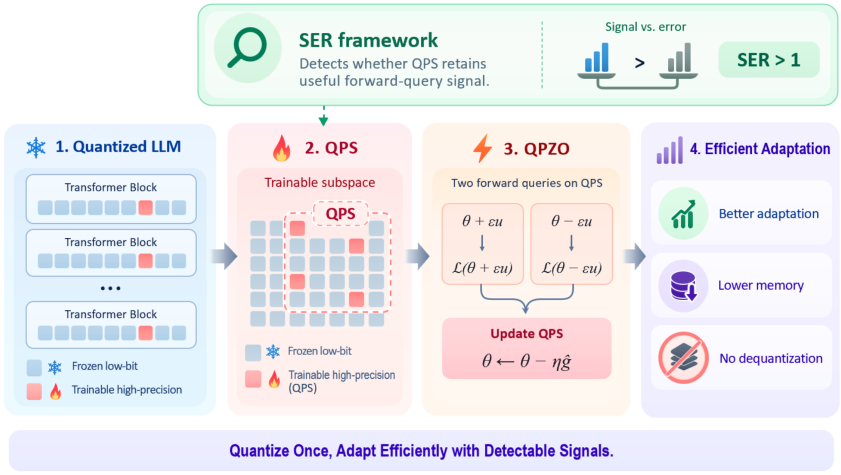


Figure 1: Overview of QPZO. QPS uses the trainable high-precision tensors retained in the quantized checkpoint. Then QPZO updates these tensors through two forward-loss queries per step, keeping the low-bit backbone fixed. This enables effective adaptation with lower training memory than full-dequantized MeZO and no dequantization for optimization. The general SER framework assesses forward-query signals in QPS, with $R_{\mathrm{se}} > 1$ certifying positive expected alignment under the stated error bound.

We further establish the feasibility of QPZO through rigorous theoretical analysis, providing guarantees for its forward-only updates. Extensive experiments demonstrate that QPZO effectively adapts quantized LLMs while preserving the compressed backbone, substantially reducing training memory while maintaining highly competitive performance. Overall, our contributions are summarized as follows:

- We identify QPS and propose QPZO for forward-only adaptation of quantized checkpoints while keeping the low-bit backbone fixed.
- We provide rigorous theoretical analysis and insights supporting the effectiveness of QPZO. SER establishes a sharp criterion for determining when finite forward queries preserve useful task directions.
- We validate QPZO across multiple models, quantizers, and tasks, demonstrating effective adaptation with favorable training-memory efficiency.

2 QPZO: QUANTIZATION-PRESERVED ZEROTH-ORDER OPTIMIZATION

In this section, we first formalize the checkpoint-native update space and the constrained optimization problem in Section 2.1. We then present the complete QPZO training procedure in Section 2.2, including its two-point update rule and forward-evaluation budget.

2.1 PROBLEM DEFINITION

Let θ_0 denote the parameters of a pretrained quantized checkpoint. Its main weight matrices are stored in low precision together with the corresponding quantization metadata, while several model tensors remain in floating point.

Let I_Q denote the coordinates of the retained floating-point tensors that are updateable. We define the **Quantization-Preserved Subspace (QPS)** as

$$S_Q = \mathrm{span}\{e_i : i \in I_Q\}, \tag{1}$$

where e_i is the coordinate basis vector associated with parameter i. Shared physical tensors are represented once in I_Q. In particular, tied input and output weights correspond to a single parameter block.

QPZO restricts adaptation to this checkpoint-native space. Let $F(\theta)$ denote the supervised training objective. The optimization problem is

$$\min_{\theta - \theta_0 \in S_Q} F(\theta). \tag{2}$$

All coordinates outside S_Q, including the packed low-bit weights and quantization metadata, remain fixed at their checkpoint values. For a mini-batch B, we denote the corresponding loss by $F_B(\theta)$.

QPS is determined directly from the checkpoint precision map and parameter sharing. It therefore requires no task-specific subspace construction and introduces no additional trainable module. Once S_Q is fixed, QPZO only needs forward-loss evaluations to optimize these coordinates.

2.2 ALGORITHM

At iteration t, QPZO samples a mini-batch B_t and a perturbation direction u_t supported on S_Q. It evaluates the same mini-batch at two symmetric perturbations of the current parameters,

$$\ell_t^+ = F_{B_t}(\theta_t + \varepsilon u_t), \qquad \ell_t^- = F_{B_t}(\theta_t - \varepsilon u_t), \tag{3}$$

where $\varepsilon > 0$ is the perturbation scale. Their difference gives the directional estimate

$$\widehat{D}_t = \frac{\ell_t^+ - \ell_t^-}{2\varepsilon}. \tag{4}$$

QPZO then updates the model along the sampled direction,

$$\theta_{t+1} = \theta_t - \eta \widehat{D}_t u_t, \tag{5}$$

where $\eta > 0$ is the learning rate. Since u_t is supported only on S_Q, every update changes only the retained floating-point tensors.

The two perturbed evaluations use the same mini-batch. After each query, the perturbed parameters are restored to θ_t before the next evaluation or update is applied. Each QPZO iteration therefore requires exactly two forward evaluations (FEs). Under a forward-evaluation budget N_{FE}, QPZO performs $N_{\mathrm{FE}}/2$ parameter updates.

Algorithm 1 presents the complete training procedure. Overall, QPZO combines quantization and ZO through a lightweight procedure, which keeps the low-bit backbone compressed and requires only two forward-loss evaluations per update. Section 3 next analyzes the two questions underlying this procedure: whether QPS contains task-relevant signal and whether finite forward queries preserve that signal well enough to produce a useful ZO update.

Algorithm 1 Quantization-Preserved Zeroth-order Optimization (QPZO)

Require: Quantized checkpoint θ_0; QPS coordinates I_Q; mini-batch loss $F_B(\theta)$; learning rate $\eta > 0$; perturbation scale $\varepsilon > 0$; forward-evaluation budget N_{FE}.
Ensure: Adapted checkpoint θ_T.
1: Keep all coordinates outside I_Q and all quantization metadata fixed.
2: Set $T \leftarrow N_{\mathrm{FE}}/2$.
3: **for** $t = 0, \ldots, T-1$ **do**
4: Sample a mini-batch B_t and an isotropic direction u_t supported on I_Q.
5: Evaluate $\ell_t^+ \leftarrow F_{B_t}(\theta_t + \varepsilon u_t)$.
6: Restore the unperturbed state θ_t.
7: Evaluate $\ell_t^- \leftarrow F_{B_t}(\theta_t - \varepsilon u_t)$.
8: Restore the unperturbed state θ_t.
9: Compute $\widehat{D}_t \leftarrow \dfrac{\ell_t^+ - \ell_t^-}{2\varepsilon}$.
10: Update $\theta_{t+1} \leftarrow \theta_t - \eta \widehat{D}_t u_t$ on I_Q.
11: **end for**
12: **return** θ_T.

3 Theoretical Analysis

In this section, we first establish the presence of task-relevant signal in QPS in Section 3.1. We then introduce SER to characterize its detectability through finite forward queries in Section 3.2, and connect positive expected alignment to a local expected-descent guarantee in Section 3.3.

3.1 Task Signal in QPS

QPZO can make progress only if the supervised objective provides a nonzero direction inside QPS. Let $g = \nabla F(\theta)$, and let P_S denote the orthogonal projector onto a subspace S. For QPS, the relevant quantity is therefore the projected gradient $P_{S_Q} g$.

The output projection provides a direct connection between task supervision and QPS. Consider an updateable untied output projection $W_{\mathrm{out}} \in S_Q$. For token-level cross entropy,

$$\nabla_{W_{\mathrm{out}}} F = \frac{1}{N_{\mathrm{tok}}} \sum_{i=1}^{N} \sum_{t \in A_i} (p_{it} - y_{it}) h_{it}^\top, \qquad \|P_{S_Q} g\|^2 \geq \|\nabla_{W_{\mathrm{out}}} F\|_F^2, \tag{6}$$

where h_{it} is the hidden representation, p_{it} is the predicted token distribution, y_{it} is the target distribution, and N_{tok} is the number of supervised tokens.

Proposition 1 (Task signal in QPS)**.** *If the aggregate gradient of an updateable untied output projection is nonzero, then*

$$P_{S_Q} g \neq 0,$$

and QPS contains a strict local descent direction. In particular,

$$\frac{\mathrm{d}}{\mathrm{d}\alpha} F(\theta - \alpha P_{S_Q} g)\Big|_{\alpha=0} = -\|P_{S_Q} g\|^2 < 0. \tag{7}$$

For tied input and output weights, both computational paths contribute to the same physical tensor, which QPS counts once. A nonzero combined gradient gives the same conclusion. The derivation is provided in Appendix B.1.

Proposition 1 establishes that task supervision can directly reach the parameters optimized by QPZO. The remaining issue is measurement: QPZO never observes $P_{S_Q} g$ directly and instead relies on the loss difference between two forward queries.

3.2 SER: Detecting Task Signal Through Forward Queries

We analyze this measurement for a general update space S, so the result applies to QPS as well as other candidate spaces. Let u be a symmetric isotropic probe supported on S, satisfying

$$\mathbb{E}[u] = 0, \qquad \mathbb{E}[uu^\top] = P_S. \tag{8}$$

The ideal directional derivative is

$$D = \langle g, u \rangle. \tag{9}$$

A forward-query protocol Π returns a measured directional value $\widehat{D}$. The protocol specifies the perturbation scale, probe distribution and normalization, mini-batch coupling, and numerical realization of the query. Its corresponding ZO estimate is

$$\widehat{g} = \widehat{D}u. \tag{10}$$

The quantity that determines whether this estimate points along a useful task direction in expectation is

$$A_{S,\Pi} := \mathbb{E}\langle g, \widehat{g} \rangle = \mathbb{E}[D\widehat{D}]. \tag{11}$$

Write the measured directional value as

$$\widehat{D} = D + E, \tag{12}$$

where E collects finite-difference, stochastic, numerical, and perturbation-realization errors. Define

$$G_S := \mathbb{E}[D^2] = \|P_S g\|^2, \qquad \sigma^2_{S,\Pi} := \mathbb{E}[E^2]. \tag{13}$$

Here, G_S measures the directional task-signal energy available in the update space, while $\sigma^2_{S,\Pi}$ measures the error introduced when that signal is accessed through the specified forward-query protocol. Their ratio defines the **Signal-to-Error Ratio (SER)**,

$$R_{\mathrm{se}}(S, \Pi; \theta) = \frac{G_S}{\sigma^2_{S,\Pi}}. \tag{14}$$

Since

$$A_{S,\Pi} = G_S + \mathbb{E}[DE] \geq G_S - \sqrt{G_S \sigma^2_{S,\Pi}}, \tag{15}$$

SER directly determines whether measurement error can erase the expected task alignment.

Theorem 1 (Sharp SER threshold for expected alignment)**.** *Let $G_S = \mathbb{E}[D^2] > 0$, and suppose the admissible measurement errors satisfy*

$$\mathbb{E}[E^2] \leq \bar{\sigma}^2.$$

Then

$$\inf_E \mathbb{E}[D\widehat{D}] = G_S - \sqrt{G_S \bar{\sigma}^2}. \tag{16}$$

Consequently,

$$\frac{G_S}{\bar{\sigma}^2} > 1$$

is the exact threshold for strictly positive expected alignment for every error satisfying the stated mean-square bound. Taking $\bar{\sigma}^2 = \sigma^2_{S,\Pi}$ gives

$$R_{\mathrm{se}}(S, \Pi; \theta) > 1 \quad \Longrightarrow \quad A_{S,\Pi} > 0. \tag{17}$$

At $R_{\mathrm{se}} = 1$, the admissible error $E = -D$ cancels the directional reading exactly. The proof is given in Appendix B.2.

SER therefore connects the task signal available in an update space with the quality of the forward measurements used to access it. For QPZO, $S = S_Q$, so $R_{\mathrm{se}} > 1$ certifies that its two-point estimator remains positively aligned with the task gradient in expectation.

3.3 From SER to Local Expected Descent

Positive expected alignment provides the first-order benefit of a QPZO update. A finite step also depends on the magnitude of the estimator and the local curvature of the objective.

Let

$$M_{2,S,\Pi} := \mathbb{E}\|\widehat{g}\|^2. \tag{18}$$

Assume that, for the update displacements considered in S, the objective satisfies the restricted smoothness condition

$$F(\theta + v) \leq F(\theta) + \langle g, v \rangle + \frac{L_S}{2}\|v\|^2. \tag{19}$$

For the ZO update

$$\theta^+ = \theta - \eta\widehat{g},$$

taking expectations gives

$$\mathbb{E}\big[F(\theta^+) - F(\theta)\big] \le -\eta A_{S,\Pi} + \frac{L_S\eta^2}{2} M_{2,S,\Pi}. \tag{20}$$

When $R_{\mathrm{se}} > 1$, Theorem 1 gives the positive alignment lower bound

$$\underline{A}_{S,\Pi} = G_S - \sqrt{G_S\sigma^2_{S,\Pi}} > 0. \tag{21}$$

Substituting this bound into Eq. (20) yields the following local guarantee.

Corollary 1 (Local expected descent). *If $R_{\mathrm{se}}(S,\Pi;\theta) > 1$ and $0 < M_{2,S,\Pi} < \infty$, then every step size satisfying*

$$0 < \eta < \frac{2\underline{A}_{S,\Pi}}{L_S M_{2,S,\Pi}} \tag{22}$$

gives

$$\mathbb{E}[F(\theta - \eta\widehat{g}) - F(\theta)] < 0. \tag{23}$$

The quadratic bound is minimized at

$$\eta_{\mathrm{rob}} = \frac{\underline{A}_{S,\Pi}}{L_S M_{2,S,\Pi}}, \tag{24}$$

for which

$$\mathbb{E}[F(\theta - \eta_{\mathrm{rob}}\widehat{g}) - F(\theta)] \le -\frac{\underline{A}^2_{S,\Pi}}{2L_S M_{2,S,\Pi}} < 0. \tag{25}$$

The proof and the multi-direction extension are provided in Appendix B.3.

This result completes the theoretical chain behind QPZO. QPS provides a task-relevant direction, SER determines whether finite forward queries preserve that direction, and the resulting alignment yields local expected descent under the stated smoothness and estimator-energy conditions. For QPZO, setting $S = S_Q$ and using the single-direction estimator in Algorithm 1 gives the update analyzed above.

4 Experiments

4.1 Main Results

Consistent gains over frozen checkpoints. Table 1 shows that QPZO outperforms Frozen in all 30 model–quantizer–task settings, covering five tasks, two quantizers, and three model families. These gains hold for both untied Qwen/Mistral and tied Gemma checkpoints, demonstrating that QPZO can reuse the trainable capacity already present in quantized checkpoints across different quantizers, architectures, and parameter-sharing structures.

Competitive performance against ZO baselines. Under the same forward-evaluation budget, QPZO outperforms QZO in most settings, showing more consistent adaptation performance across models and downstream tasks. It also remains competitive with full-dequantized MeZO, achieving better results in 16 of the 30 settings. Importantly, QPZO achieves this performance without maintaining a fully dequantized parameter space for optimization, updating only a substantially smaller set of trainable parameters.

Lower training memory. Figure 2 shows that QPZO uses substantially less peak training memory than full-dequantized MeZO across all six model–quantizer settings. On Qwen and Mistral, QPZO requires only 0.12–0.54 GiB more than QZO, a specialized baseline that updates only quantization scales and does not retain full-precision shadow weights. Thus, on these models, QPZO remains close to QZO in memory usage despite directly updating the retained floating-point model tensors. These results demonstrate a strong memory advantage while providing a richer checkpoint-native parameter space for adaptation.

Table 1: Classification tasks report accuracy and SQuAD reports F1. Best results within each setting are shown in bold.

Model	Method	Classification ↑ SST-2	RTE	WiC	BoolQ	Generation ↑ SQuAD F1
Qwen2.5-7B GPTQ	Frozen	90.94	76.53	56.74	81.80	56.21
	QZO	94.72	**86.28**	55.96	84.80	**87.55**
	MeZO	93.69	**86.28**	64.42	84.28	87.00
	QPZO	**94.95**	**86.28**	**71.79**	**85.87**	83.69
Qwen2.5-7B AWQ	Frozen	90.94	79.78	66.93	79.69	57.85
	QZO	91.51	83.75	57.52	79.91	**87.37**
	MeZO	92.32	88.81	64.89	82.72	87.34
	QPZO	**95.07**	**89.89**	**69.59**	**83.91**	84.82
Mistral-7B GPTQ	Frozen	89.79	75.09	53.61	82.17	75.73
	QZO	84.40	58.84	51.25	64.10	79.85
	MeZO	80.28	**76.53**	**62.54**	85.14	**87.87**
	QPZO	**92.55**	75.45	60.97	**85.23**	81.42
Mistral-7B AWQ	Frozen	90.60	75.09	56.90	81.25	64.82
	QZO	54.59	46.21	53.45	58.53	83.91
	MeZO	77.29	50.54	57.99	59.11	**86.54**
	QPZO	**93.00**	**76.17**	**62.23**	**83.85**	80.99
Gemma-2-9B GPTQ	Frozen	82.91	84.12	62.23	79.88	70.98
	QZO	**95.53**	**85.20**	52.04	63.06	91.42
	MeZO	94.72	84.84	**69.59**	**88.20**	**91.49**
	QPZO	95.30	84.48	65.52	83.20	90.94
Gemma-2-9B AWQ	Frozen	84.98	82.31	67.08	79.11	71.72
	QZO	94.50	81.95	60.34	85.63	88.85
	MeZO	94.72	**85.92**	63.64	**87.43**	**91.56**
	QPZO	**95.18**	83.39	**68.18**	81.70	90.34

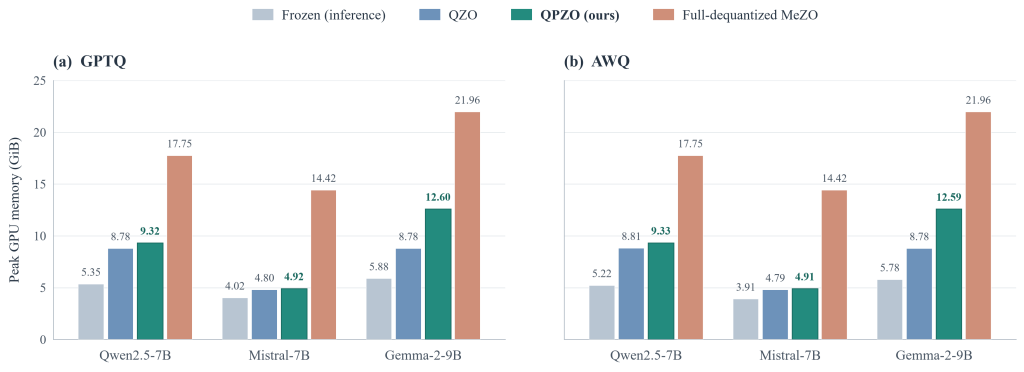


Figure 2: Peak allocated GPU memory on SST-2 for (a) GPTQ and (b) AWQ checkpoints. Frozen reports inference memory, whereas QZO, QPZO, and full-dequantized MeZO report training memory. QZO updates only quantization scales and does not retain full-precision shadow weights. QPZO uses substantially less training memory than full-dequantized MeZO across all six model–quantizer settings.

4.2 Ablation Study

4.2.1 Why Does QPS Work for Zeroth-Order Updates?

We first examine why QPS provides an effective update space for zeroth-order optimization. We examine this assumption directly on Qwen2.5-7B GPTQ/SST-2 by comparing the following candidate update spaces: (1) **QPS**: QPS used by QPZO; (2) **Output head**: only the untied output-projection weights; (3) **QPS+LoRA**: QPS together with rank-16 LoRA updates on low-precision projections; (4) **Low-precision LoRA**: only the rank-16 LoRA on low-precision projections; (5) **Dequantized low-precision weights**: the continuous weights obtained by dequantizing the low-precision projections; (6) **Matched random**: five random spaces, each approximately dimension-matched to QPS; and (7) **Norm-only**: only the RMSNorm weight vectors. For each candidate, we use the finite-sample SER certificate in Appendix B.4 to evaluate whether its directional task signal remains detectable through forward queries, and further test whether the resulting local update produces a validated loss decrease.

Table 2: SER ablation on Qwen2.5-7B GPTQ/SST-2 across five runs. Panel (a) compares the certified SER ratio and local-step outcomes across update spaces. Panel (b) follows QPS from detectable directional signal to an independently accepted loss decrease. Values report mean from five seeds. Validated decreases denote candidate local steps that pass the SER criterion and are subsequently accepted by an independent loss-decrease evaluation. Post-hoc decreases denote negative loss changes identified by the same empirical evaluation in the diagnostic-only branch despite not passing the SER criterion. The reported M_2, η_{search}, and loss-change upper bound are offline diagnostic quantities.

(a) Adaptive SER and candidate update space outcomes

Subspace–protocol pair	SER ratio mean	SER passes	Validated decreases	Post-hoc decreases
QPS (T)	3.021	5/5	5/5	0/5
Output head (H)	3.114	5/5	4/5	0/5
QPS+LoRA-r16 (T)	2.011	5/5	0/5	0/5
Low-precision LoRA-r16 (T)	0.904	0/5	0/5	3/5
Dequantized low precision (T)	0.015	0/5	0/5	2/5
Matched random 000 (T)	0.950	1/5	1/5	3/5
Matched random 001 (T)	0.959	0/5	0/5	4/5
Matched random 002 (T)	0.915	1/5	1/5	3/5
Matched random 003 (T)	0.749	1/5	0/5	3/5
Matched random 004 (T)	0.873	0/5	0/5	3/5
Pooled matched random	0.889	3/25	2/25	16/25
Norm only (T)	abstention	0/5	0/5	0/5

(b) Same-instance detectability-to-descent closure for QPS

Metric	Mean	Metric	Mean
Certified SER ratio $\underline{G}/\overline{\sigma}^2$	3.021	Alignment bound $\underline{A}$ ($\times 10^3$)	5.802
Estimator-energy bound $\overline{M}_2$ ($\times 10^{13}$)	4.191	Progress score $\underline{Z}$ ($\times 10^{-7}$)	8.134
Selected step η_{search} ($\times 10^{-5}$)	1.053	Loss-change bound $\overline{\Delta F}_{\text{acc}}$	-0.4467

QPS gives the most complete and consistent result among all candidate update spaces in Table 2. It passes the SER criterion in all five runs and is the only candidate that also produces a validated decrease in all five. Its certified SER ratio remains between 2.736 and 3.441, while the dimension-matched random spaces pass only 3 of 25 pooled runs and lie only slightly above the threshold when they do pass. This gap shows that the advantage of QPS comes from the task-relevant structure already preserved in the checkpoint, rather than dimensionality alone. Panel (b) further completes the chain for QPS: a certified SER above one leads to a positive alignment bound and progress score, followed by a negative independently evaluated loss-change bound. Some uncertified spaces still

exhibit post-hoc decreases, which is consistent with SER serving as a sufficient certificate rather than a necessary condition for every useful update. Overall, QPS is the only tested space that consistently connects detectable forward-query signal, positive local progress, and validated descent across all runs.

4.2.2 Is the Output Head Sufficient

Table 3: Analytical-head comparison on untied checkpoints. Each untied Qwen/Mistral model–quantizer cell reports single-run scores. SQuAD reports F1, the remaining tasks report accuracy.

Model	Method	Classification				Generation
		SST-2	RTE	WiC	BoolQ	SQuAD
Qwen2.5-7B GPTQ	Exact streaming head	94.04	**86.64**	**71.79**	85.75	55.34
	Head-only ZO	**94.95**	86.28	**71.79**	85.66	62.26
	QPZO	**94.95**	86.28	**71.79**	**85.87**	**83.69**
Qwen2.5-7B AWQ	Exact streaming head	93.69	89.17	**70.53**	84.25	60.39
	Head-only ZO	**95.18**	89.53	69.75	**85.17**	77.75
	QPZO	95.07	**89.89**	69.59	83.91	**84.82**
Mistral-7B GPTQ	Exact streaming head	90.94	73.29	**63.17**	85.57	80.68
	Head-only ZO	91.06	73.29	61.29	**85.72**	81.14
	QPZO	**92.55**	**75.45**	60.97	85.23	**81.42**
Mistral-7B AWQ	Exact streaming head	91.51	76.53	60.82	**85.44**	80.40
	Head-only ZO	91.51	**76.90**	59.40	85.11	**81.46**
	QPZO	**93.00**	76.17	**62.23**	83.85	80.99

Although the output head provides the most direct source of task signal in QPS, it may not capture the full adaptation capacity of the retained parameter space. To assess the contribution of the remaining QPS tensors, we compare three strategies on untied Qwen and Mistral checkpoints: (1) Exact streaming-head adaptation updates only the output projection using its analytical gradient, while keeping all other parameters fixed. (2) Head-only ZO updates the same output projection using two-point forward-loss queries, while keeping all other parameters fixed. (3) QPZO applies two-point ZO updates to the complete QPS.

Table 3 shows the performance of these three strategies. QPZO outperforms the exact streaming-head baseline on all four SST-2 settings and all four SQuAD settings. The difference is especially large on Qwen SQuAD, where QPZO improves over head-only ZO from 62.26 to 83.69 F1 under GPTQ and from 77.75 to 84.82 under AWQ. These results show that the output head supplies an important task-dependent signal, while the other retained floating-point tensors in QPS contribute additional useful directions. The effectiveness of QPZO therefore comes from the complete checkpoint-native update space, rather than from the output projection alone.

5 Conclusion and Limitations

We introduce QPZO, which directly adapts the QPS already available in quantized checkpoints. By reusing these preserved high-precision tensors, QPZO requires no additional trainable parameters or task-specific subspace search, keeps the low-bit backbone compressed, avoids dequantization, and achieves strong downstream performance with lower training memory across multiple model families, quantization methods, and NLP tasks. SER provides a general framework for evaluating whether forward queries preserve useful directional signals across different update spaces and protocols. Our current evaluation focuses on 4-bit GPTQ/AWQ language models and NLP tasks and future work will extend QPZO to lower-bit quantization, broader model families, and more efficient SER estimation.

AI USE STATEMENT

Generative AI tools assisted with language polishing, literature searches, research ideation, manuscript drafting, and developing and checking mathematical proofs. The authors independently verified all AI-assisted content and take full responsibility for the final paper.

ETHICS STATEMENT

We have manually re-examined the datasets used in this study to ensure that they are free from potential discrimination, human rights violations, bias, exploitation, and other ethical concerns.

REPRODUCIBILITY STATEMENT

To facilitate reproducibility, we provide the code used in our experiments in the supplementary material.

REFERENCES

Yiming Chen, Yuan Zhang, Liyuan Cao, Kun Yuan, and Zaiwen Wen. Enhancing zeroth-order fine-tuning for language models with low-rank structures. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=9BiVepgmWW.

Christopher Clark, Kenton Lee, Ming-Wei Chang, Tom Kwiatkowski, Michael Collins, and Kristina Toutanova. BoolQ: Exploring the surprising difficulty of natural yes/no questions. In Jill Burstein, Christy Doran, and Thamar Solorio (eds.), *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 2924–2936, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1300. URL https://aclanthology.org/N19-1300/.

Ido Dagan, Oren Glickman, and Bernardo Magnini. The PASCAL recognising textual entailment challenge. In Joaquin Quiñonero Candela, Ido Dagan, Bernardo Magnini, and Florence d'Alché-Buc (eds.), *Machine Learning Challenges, Evaluating Predictive Uncertainty, Visual Object Classification and Recognizing Textual Entailment, First PASCAL Machine Learning Challenges Workshop, MLCW 2005, Southampton, UK, April 11-13, 2005, Revised Selected Papers*, volume 3944 of *Lecture Notes in Computer Science*, pp. 177–190. Springer, 2005. doi: 10.1007/11736790_9. URL https://doi.org/10.1007/11736790_9.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. QLoRA: Efficient finetuning of quantized LLMs. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 10088–10115. Curran Associates, Inc., 2023. doi: 10.52202/075280-0441. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/1feb87871436031bdc0f2beaa62a049b-Paper-Conference.pdf.

Olivier Devolder, François Glineur, and Yurii E. Nesterov. First-order methods of smooth convex optimization with inexact oracle. *Math. Program.*, 146(1-2):37–75, 2014. doi: 10.1007/S10107-013-0677-5. URL https://doi.org/10.1007/s10107-013-0677-5.

John C. Duchi, Michael I. Jordan, Martin J. Wainwright, and Andre Wibisono. Optimal rates for zero-order convex optimization: The power of two function evaluations. *IEEE Trans. Inf. Theory*, 61(5):2788–2806, 2015. doi: 10.1109/TIT.2015.2409256. URL https://doi.org/10.1109/TIT.2015.2409256.

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.

Tanmay Gautam, Youngsuk Park, Hao Zhou, Parameswaran Raman, and Wooseok Ha. Variance-reduced zeroth-order methods for fine-tuning language models. In Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp (eds.), *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 15180–15208. PMLR, 21–27 Jul 2024. URL https://proceedings.mlr.press/v235/gautam24a.html.

Saeed Ghadimi and Guanghui Lan. Stochastic first- and zeroth-order methods for nonconvex stochastic programming. *SIAM J. Optim.*, 23(4):2341–2368, 2013. doi: 10.1137/120880811. URL https://doi.org/10.1137/120880811.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. Mistral 7b, 2023. URL https://arxiv.org/abs/2310.06825.

Yixiao Li, Yifan Yu, Chen Liang, Nikos Karampatziakis, Pengcheng He, Weizhu Chen, and Tuo Zhao. Loftq: LoRA-fine-tuning-aware quantization for large language models. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=LzPWWPAdY4.

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 P. Gibbons, G. Pekhimenko, and C. De Sa (eds.), *Proceedings of Machine Learning and Systems*, volume 6, pp. 87–100, 2024. URL https://proceedings.mlsys.org/paper_files/paper/2024/file/42a452cbafa9dd64e9ba4aa95cc1ef21-Paper-Conference.pdf.

Yong Liu, Zirui Zhu, Chaoyu Gong, Minhao Cheng, Cho-Jui Hsieh, and Yang You. Sparse MeZO: Less parameters for better performance in zeroth-order LLM fine-tuning. In D. Belgrave, C. Zhang, H. Lin, R. Pascanu, P. Koniusz, M. Ghassemi, and N. Chen (eds.), *Advances in Neural Information Processing Systems*, volume 38, Main Conference, pp. 21139–21169. Curran Associates, Inc., 2025. doi: 10.52202/085713-0712. URL https://proceedings.neurips.cc/paper_files/paper/2025/file/1e5c2efbddc02c1d971e2f19ccdb07d0-Paper-Conference.pdf.

Sadhika Malladi, Tianyu Gao, Eshaan Nichani, Alex Damian, Jason D Lee, Danqi Chen, and Sanjeev Arora. Fine-tuning language models with just forward passes. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 53038–53075. Curran Associates, Inc., 2023. doi: 10.52202/075280-2308. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/a627810151be4d13f907ac898ff7e948-Paper-Conference.pdf.

Mohammad Taher Pilehvar and Jose Camacho-Collados. WiC: the word-in-context dataset for evaluating context-sensitive meaning representations. In Jill Burstein, Christy Doran, and Thamar Solorio (eds.), *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 1267–1273, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1128. URL https://aclanthology.org/N19-1128/.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. SQuAD: 100,000+ questions for machine comprehension of text. In Jian Su, Kevin Duh, and Xavier Carreras (eds.), *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pp. 2383–2392, Austin, Texas, November 2016. Association for Computational Linguistics. doi: 10.18653/v1/D16-1264. URL https://aclanthology.org/D16-1264/.

Sifeng SHANG, JIAYI ZHOU, Chenyu Lin, Minxian Li, and Kaiyang Zhou. Fine-tuning quantized neural networks with zeroth-order optimization. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=mGeeRFToaW.

Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D. Manning, Andrew Ng, and Christopher Potts. Recursive deep models for semantic compositionality over a sentiment treebank. In David Yarowsky, Timothy Baldwin, Anna Korhonen, Karen Livescu, and Steven Bethard (eds.), *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pp. 1631–1642, Seattle, Washington, USA, October 2013. Association for Computational Linguistics. URL https://aclanthology.org/D13-1170/.

J.C. Spall. Multivariate stochastic approximation using a simultaneous perturbation gradient approximation. *IEEE Transactions on Automatic Control*, 37(3):332–341, 1992.

Gemma Team, Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé, Johan Ferret, Peter Liu, Pouya Tafti, Abe Friesen, Michelle Casbon, Sabela Ramos, Ravin Kumar, Charline Le Lan, Sammy Jerome, Anton Tsitsulin, Nino Vieillard, Piotr Stanczyk, Sertan Girgin, Nikola Momchev, Matt Hoffman, Shantanu Thakoor, Jean-Bastien Grill, Behnam Neyshabur, Olivier Bachem, Alanna Walton, Aliaksei Severyn, Alicia Parrish, Aliya Ahmad, Allen Hutchison, Alvin Abdagic, Amanda Carl, Amy Shen, Andy Brock, Andy Coenen, Anthony Laforge, Antonia Paterson, Ben Bastian, Bilal Piot, Bo Wu, Brandon Royal, Charlie Chen, Chintu Kumar, Chris Perry, Chris Welty, Christopher A. Choquette-Choo, Danila Sinopalnikov, David Weinberger, Dimple Vijaykumar, Dominika Rogozińska, Dustin Herbison, Elisa Bandy, Emma Wang, Eric Noland, Erica Moreira, Evan Senter, Evgenii Eltyshev, Francesco Visin, Gabriel Rasskin, Gary Wei, Glenn Cameron, Gus Martins, Hadi Hashemi, Hanna Klimczak-Plucińska, Harleen Batra, Harsh Dhand, Ivan Nardini, Jacinda Mein, Jack Zhou, James Svensson, Jeff Stanway, Jetha Chan, Jin Peng Zhou, Joana Carrasqueira, Joana Iljazi, Jocelyn Becker, Joe Fernandez, Joost van Amersfoort, Josh Gordon, Josh Lipschultz, Josh Newlan, Ju yeong Ji, Kareem Mohamed, Kartikeya Badola, Kat Black, Katie Millican, Keelin McDonell, Kelvin Nguyen, Kiranbir Sodhia, Kish Greene, Lars Lowe Sjoesund, Lauren Usui, Laurent Sifre, Lena Heuermann, Leticia Lago, Lilly McNealus, Livio Baldini Soares, Logan Kilpatrick, Lucas Dixon, Luciano Martins, Machel Reid, Manvinder Singh, Mark Iverson, Martin Görner, Mat Velloso, Mateo Wirth, Matt Davidow, Matt Miller, Matthew Rahtz, Matthew Watson, Meg Risdal, Mehran Kazemi, Michael Moynihan, Ming Zhang, Minsuk Kahng, Minwoo Park, Mofi Rahman, Mohit Khatwani, Natalie Dao, Nenshad Bardoliwalla, Nesh Devanathan, Neta Dumai, Nilay Chauhan, Oscar Wahltinez, Pankil Botarda, Parker Barnes, Paul Barham, Paul Michel, Pengchong Jin, Petko Georgiev, Phil Culliton, Pradeep Kuppala, Ramona Comanescu, Ramona Merhej, Reena Jana, Reza Ardeshir Rokni, Rishabh Agarwal, Ryan Mullins, Samaneh Saadat, Sara Mc Carthy, Sarah Cogan, Sarah Perrin, Sébastien M. R. Arnold, Sebastian Krause, Shengyang Dai, Shruti Garg, Shruti Sheth, Sue Ronstrom, Susan Chan, Timothy Jordan, Ting Yu, Tom Eccles, Tom Hennigan, Tomas Kocisky, Tulsee Doshi, Vihan Jain, Vikas Yadav, Vilobh Meshram, Vishal Dharmadhikari, Warren Barkley, Wei Wei, Wenming Ye, Woohyun Han, Woosuk Kwon, Xiang Xu, Zhe Shen, Zhitao Gong, Zichuan Wei, Victor Cotruta, Phoebe Kirk, Anand Rao, Minh Giang, Ludovic Peran, Tris Warkentin, Eli Collins, Joelle Barral, Zoubin Ghahramani, Raia Hadsell, D. Sculley, Jeanine Banks, Anca Dragan, Slav Petrov, Oriol Vinyals, Jeff Dean, Demis Hassabis, Koray Kavukcuoglu, Clement Farabet, Elena Buchatskaya, Sebastian Borgeaud, Noah Fiedel, Armand Joulin, Kathleen Kenealy, Robert Dadashi, and Alek Andreev. Gemma 2: Improving open language models at a practical size, 2024. URL https://arxiv.org/abs/2408.00118.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In Tal Linzen, Grzegorz Chrupała, and Afra Alishahi (eds.), *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pp. 353–355, Brussels, Belgium, November 2018. Association for Computational Linguistics. doi: 10.18653/v1/W18-5446. URL https://aclanthology.org/W18-5446/.

Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. SuperGLUE: A stickier benchmark for general-purpose language understanding systems. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL https://proceedings.neurips.cc/paper_files/paper/2019/file/4496bf24afe7fab6f046bf4923da8de6-Paper.pdf.

Yuhui Xu, Lingxi Xie, Xiaotao Gu, Xin Chen, Heng Chang, Hengheng Zhang, Zhengsu Chen, XIAOPENG ZHANG, and Qi Tian. QA-loRA: Quantization-aware low-rank adaptation of large language models. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=WvFoJccpo8.

Qwen: An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tianyi Tang, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. Qwen2.5 technical report, 2025. URL https://arxiv.org/abs/2412.15115.

Ziming Yu, Pan Zhou, Sike Wang, Jia Li, Mi Tian, and Hua Huang. Zeroth-order fine-tuning of LLMs in random subspaces. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 4475–4485, October 2025.

Yihua Zhang, Pingzhi Li, Junyuan Hong, Jiaxiang Li, Yimeng Zhang, Wenqing Zheng, Pin-Yu Chen, Jason D. Lee, Wotao Yin, Mingyi Hong, Zhangyang Wang, Sijia Liu, and Tianlong Chen. Revisiting zeroth-order optimization for memory-efficient LLM fine-tuning: A benchmark. In Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp (eds.), *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 59173–59190. PMLR, 21–27 Jul 2024. URL https://proceedings.mlr.press/v235/zhang24ad.html.

Jiajun Zhou, Yifan Yang, Kai Zhen, Ziyue Liu, Yequan Zhao, Ershad Banijamali, Athanasios Mouchtaris, Ngai Wong, and Zheng Zhang. QuZO: Quantized zeroth-order fine-tuning for large language models. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng (eds.), *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 5341–5359, Suzhou, China, November 2025. Association for Computational Linguistics. ISBN 979-8-89176-332-6. doi: 10.18653/v1/2025.emnlp-main.271. URL https://aclanthology.org/2025.emnlp-main.271/.

A Related Work

Zeroth-order LLM fine-tuning and subspace perturbations. Zeroth-order (ZO) optimization estimates update directions from function values, removing the need to backpropagate through the model (Spall, 1992). MeZO brought this idea to large language models by performing two-point updates in place, enabling fine-tuning with an inference-level memory footprint (Malladi et al., 2023). Subsequent work has improved the practicality of ZO fine-tuning through broader benchmarking and variance reduction (Zhang et al., 2024; Gautam et al., 2024). A second line of work concerns where the perturbation should be applied. Sparse MeZO updates a selected subset of weights, LOZO exploits low-rank structure, and SubZero performs ZO fine-tuning in random low-dimensional subspaces (Liu et al., 2025; Chen et al., 2025; Yu et al., 2025). Together, these studies show that the choice of update space strongly affects ZO efficiency. Our method extends this idea to quantized checkpoints by reading an update space directly from the model, which is the floating-point tensors already retained alongside the low-bit backbone. This removes a separate task-specific subspace-construction stage and gives ZO a set of parameters that can be perturbed continuously.

Adaptation of quantized language models. Post-training quantization methods such as GPTQ (Frantar et al., 2023) and AWQ (Lin et al., 2024) make large models substantially cheaper to store and serve by compressing their main weight matrices to low precision (Frantar et al., 2023; Lin et al., 2024). Several methods combine quantization with parameter-efficient fine-tuning. For instance, QLoRA keeps the quantized backbone frozen and backpropagates into low-rank adapters, while LoftQ and QA-LoRA further coordinate quantization with low-rank adaptation (Dettmers et al., 2023; Li et al., 2024; Xu et al., 2024). Recent work builds on the idea of zeroth-order optimization. For example, QuZO performs low-precision ZO fine-tuning with optimized stochastic rounding, and QZO perturbs continuous quantization scales while keeping the discrete weight codes fixed (Zhou et al., 2025; SHANG et al., 2026). These approaches establish that quantized models can support effective

task adaptation through several trainable parameterizations. QPZO uses the trainable capacity that is already present in the checkpoint. It keeps the low-bit weights and quantization metadata fixed and updates the retained floating-point tensors directly. As a result, low-bit weights do not need to be dequantized for optimization or requantized after each update, and no additional adapter or trainable quantization parameterization is required.

Reliable forward-query signals. Classical ZO analyses study how finite function queries, stochastic noise, and parameter dimension affect the quality and convergence of random-direction estimators (Spall, 1992; Ghadimi & Lan, 2013; Duchi et al., 2015). Related inexact-oracle work likewise shows that optimization behavior depends on the accuracy of the information supplied to the optimizer (Devolder et al., 2014). In a quantized implementation, a two-point ZO step relies on a small loss difference that can be affected by finite-difference error, mini-batch variation, numerical rounding, and imperfect realization of the requested perturbation. Our SER analysis turns this measurement question into a simple criterion by comparing directional signal energy with mean-squared measurement error. We establish $R_{\mathrm{se}} > 1$ as the sharp threshold for robust positive expected alignment under the stated mean-square error bound. This gives the two pieces of QPZO complementary roles. QPS supplies a ready-to-use update space from the checkpoint, and SER determines whether the chosen forward-query protocol can reliably use the task signal available in that space.

B PROOFS

B.1 PROOF OF PROPOSITION 1

Consider the updateable untied output projection W_{out} in Section 3.1. For one supervised token, let h denote its hidden state, y its target distribution, and

$$z = W_{\mathrm{out}} h, \qquad p = \mathrm{softmax}(z), \qquad \ell = -y^\top \log p. \tag{26}$$

For softmax cross entropy,

$$\nabla_z \ell = p - y. \tag{27}$$

Since the output head is untied, h is fixed when differentiating with respect to W_{out}. The chain rule therefore gives

$$\nabla_{W_{\mathrm{out}}} \ell = (p - y) h^\top. \tag{28}$$

Let A_i denote the supervised token positions in example i, and let

$$N_{\mathrm{tok}} = \sum_{i=1}^{N} |A_i| > 0. \tag{29}$$

Averaging over all supervised tokens gives

$$\nabla_{W_{\mathrm{out}}} F = \frac{1}{N_{\mathrm{tok}}} \sum_{i=1}^{N} \sum_{t \in A_i} (p_{it} - y_{it}) h_{it}^\top.$$

Because the output-head coordinates form one physical block of S_Q,

$$\|P_{S_Q} g\|^2 \geq \|\nabla_{W_{\mathrm{out}}} F\|_F^2.$$

Hence, whenever the aggregate output-head gradient is nonzero,

$$P_{S_Q} g \neq 0.$$

Moreover,

$$\frac{\mathrm{d}}{\mathrm{d}\alpha} F(\theta - \alpha P_{S_Q} g)\Big|_{\alpha=0} = -\|P_{S_Q} g\|^2 < 0,$$

so QPS contains a strict local descent direction.

Under the isotropic probe normalization

$$\mathbb{E}[uu^\top] = P_{S_Q},$$

the corresponding directional signal energy is

$$G_{S_Q} = \mathbb{E}[(g^\top u)^2] = g^\top P_{S_Q} g = \|P_{S_Q} g\|^2 \geq \|\nabla_{W_{\text{out}}} F\|_F^2. \tag{30}$$

Thus a nonzero aggregate output-head gradient also implies $G_{S_Q} > 0$.

For tied input and output weights, the embedding and output paths contribute to the same physical tensor. QPS counts this tensor once and uses its complete gradient. A nonzero combined gradient therefore yields the same conclusion.

B.2 Proof of Theorem 1

Let

$$0 < G_S = \mathbb{E}[D^2] < \infty$$

and let the admissible measurement-error class be

$$\mathcal{E}(\bar{\sigma}) = \left\{ E \in L_2 : \mathbb{E}[E^2] \leq \bar{\sigma}^2 \right\}. \tag{31}$$

The error may be biased and may depend on D. For every $E \in \mathcal{E}(\bar{\sigma})$,

$$\begin{aligned} \mathbb{E}[D\widehat{D}] &= \mathbb{E}[D(D+E)] \\ &= G_S + \mathbb{E}[DE] \\ &\geq G_S - \sqrt{\mathbb{E}[D^2]\mathbb{E}[E^2]} \\ &\geq G_S - \sqrt{G_S \bar{\sigma}^2}, \end{aligned} \tag{32}$$

where the first inequality follows from Cauchy–Schwarz.

The bound is attained by

$$E^\star = -\frac{\bar{\sigma}}{\sqrt{G_S}} D. \tag{33}$$

Indeed,

$$\mathbb{E}[(E^\star)^2] = \bar{\sigma}^2,$$

so $E^\star$ is admissible, and

$$\mathbb{E}[D(D+E^\star)] = G_S - \sqrt{G_S \bar{\sigma}^2}.$$

Therefore,

$$\inf_{E \in \mathcal{E}(\bar{\sigma})} \mathbb{E}[D(D+E)] = G_S - \sqrt{G_S \bar{\sigma}^2}.$$

This quantity is strictly positive exactly when

$$G_S > \bar{\sigma}^2.$$

For $\bar{\sigma}^2 > 0$,

$$\frac{G_S}{\bar{\sigma}^2} > 1 \quad \Longleftrightarrow \quad \inf_{E \in \mathcal{E}(\bar{\sigma})} \mathbb{E}[D\widehat{D}] > 0. \tag{34}$$

Taking

$$\bar{\sigma}^2 = \sigma_{S,\Pi}^2 = \mathbb{E}[E^2]$$

recovers the SER criterion

$$R_{\text{se}}(S, \Pi; \theta) > 1.$$

The same construction gives the boundary cases. At $G_S = \bar{\sigma}^2 > 0$, the admissible error $E^\star = -D$ removes the directional reading and yields zero expected alignment. When $0 < G_S < \bar{\sigma}^2$, the worst-case alignment is negative. If $\bar{\sigma}^2 = 0 < G_S$, the error vanishes almost surely and the alignment equals $G_S > 0$.

This proves the sharp threshold stated in Theorem 1.

B.3 Proof of Corollary 1 and Multi-Direction Extension

We first state the argument for a general m-direction estimator. QPZO corresponds to the special case $m = 1$.

For m directional measurements, define

$$\widehat{g}_{S,m} = \frac{1}{m}\sum_{j=1}^{m} \widehat{D}_j u_j, \qquad D_j = \langle g, u_j \rangle, \tag{35}$$

where each pair $(u_j, \widehat{D}_j)$ follows the same marginal query protocol. Define

$$A_{S,m} = \mathbb{E}\langle g, \widehat{g}_{S,m}\rangle, \qquad M_{2,S,m} = \mathbb{E}\|\widehat{g}_{S,m}\|^2. \tag{36}$$

By linearity,

$$\begin{aligned} A_{S,m} &= \frac{1}{m}\sum_{j=1}^{m} \mathbb{E}[\widehat{D}_j \langle g, u_j\rangle] \\ &= \frac{1}{m}\sum_{j=1}^{m} \mathbb{E}[D_j \widehat{D}_j] \\ &= \mathbb{E}[D\widehat{D}]. \end{aligned} \tag{37}$$

This identity does not require independence across the sampled pairs.

Applying Theorem 1 to the marginal protocol gives

$$A_{S,m} \geq \underline{A}_S := G_S - \sqrt{G_S \sigma^2_{S,\Pi}}. \tag{38}$$

Therefore,

$$R_{\mathrm{se}}(S, \Pi; \theta) > 1 \quad \Longrightarrow \quad \underline{A}_S > 0.$$

Now assume

$$0 < M_{2,S,m} < \infty,$$

and let $V_S \subseteq S$ be a set of admissible displacements such that, for some $0 < L_S < \infty$,

$$F(\theta + v) \leq F(\theta) + \langle g, v\rangle + \frac{L_S}{2}\|v\|^2, \qquad v \in V_S. \tag{39}$$

Consider

$$\theta^+ = \theta - \eta \widehat{g}_{S,m},$$

where $\eta > 0$ is fixed before sampling the estimator and

$$-\eta \widehat{g}_{S,m} \in V_S$$

almost surely.

Substituting the update into the smoothness inequality gives

$$F(\theta^+) - F(\theta) \leq -\eta \langle g, \widehat{g}_{S,m}\rangle + \frac{L_S \eta^2}{2}\|\widehat{g}_{S,m}\|^2. \tag{40}$$

Taking expectations,

$$\begin{aligned} \mathbb{E}[F(\theta^+) - F(\theta)] &\leq -\eta A_{S,m} + \frac{L_S \eta^2}{2} M_{2,S,m} \\ &\leq -\eta \underline{A}_S + \frac{L_S \eta^2}{2} M_{2,S,m}. \end{aligned} \tag{41}$$

When $R_{\mathrm{se}} > 1$, we have $\underline{A}_S > 0$. Hence every step size satisfying

$$0 < \eta < \frac{2\underline{A}_S}{L_S M_{2,S,m}} \tag{42}$$

gives a strictly negative upper bound on the expected loss change.

The quadratic upper bound is minimized at

$$\eta_{\mathrm{rob}} = \frac{\underline{A}_S}{L_S M_{2,S,m}}. \tag{43}$$

When this step also satisfies the displacement condition,

$$\mathbb{E}\big[F(\theta - \eta_{\mathrm{rob}}\widehat{g}_{S,m}) - F(\theta)\big] \le -\frac{\underline{A}_S^2}{2L_S M_{2,S,m}} < 0. \tag{44}$$

This proves the local expected-descent result.

More generally, suppose valid bounds

$$\underline{A} \le A_{S,m}, \qquad 0 < M_{2,S,m} \le \overline{M}_2 < \infty \tag{45}$$

are fixed before drawing the update estimator. Define

$$Z = \frac{[\underline{A}]_+^2}{\overline{M}_2}, \qquad [a]_+ = \max\{a, 0\}. \tag{46}$$

For $\underline{A} > 0$, choosing

$$\eta = \frac{\underline{A}}{L_S \overline{M}_2} \tag{47}$$

gives

$$\mathbb{E}[F(\theta - \eta\widehat{g}_{S,m}) - F(\theta)] \le -\frac{Z}{2L_S}. \tag{48}$$

This is the progress quantity reported in the SER ablation.

For QPZO,

$$S = S_Q, \qquad m = 1,$$

and $\widehat{g}_{S,1} = \widehat{D}u$ is exactly the two-point estimator used in Algorithm 1. Thus the single-direction result in Section 3.3 follows directly from the general argument above.

B.4 Finite-Sample SER Certificate

The SER in Section 3.2 is defined through the population quantities $G_S = \mathbb{E}[D^2]$ and $\sigma_{S,\Pi}^2 = \mathbb{E}[E^2]$. These quantities are useful for analysis, but they are not directly observed from a finite number of forward queries. We therefore use a finite-sample certificate that estimates the signal and error energies separately and checks whether the signal still exceeds the error after accounting for estimation uncertainty.

Replicated forward measurements. For each sampled direction u_j, let

$$X_j = D_j + E_j \tag{49}$$

denote the directional reading produced by the evaluated query protocol, where

$$D_j = \langle \nabla F(\theta), u_j \rangle.$$

We additionally obtain K independent reference readings

$$Y_{j,r} = D_j + \zeta_{j,r}, \qquad r = 1, \dots, K. \tag{50}$$

The reference error $\zeta_{j,r}$ is conditionally centered,

$$\mathbb{E}[\zeta_{j,r} \mid u_j] = 0, \tag{51}$$

and the reference replicates are conditionally independent given u_j. They are also independent of the candidate error E_j conditional on the sampled direction. The candidate error itself may be biased.

The repeated references allow the signal and measurement error to be separated. For each direction, define

$$Z_j^G = \frac{1}{K(K-1)} \sum_{r \neq s} Y_{j,r} Y_{j,s}, \tag{52}$$

and

$$Z_j^\sigma = \frac{1}{K(K-1)} \sum_{r \neq s} (X_j - Y_{j,r})(X_j - Y_{j,s}). \tag{53}$$

For $r \neq s$, conditional independence gives

$$\mathbb{E}[Y_{j,r} Y_{j,s} \mid u_j] = D_j^2, \tag{54}$$

while $X_j - Y_{j,r} = E_j - \zeta_{j,r}$ gives

$$\mathbb{E}[(X_j - Y_{j,r})(X_j - Y_{j,s}) \mid u_j, X_j] = E_j^2. \tag{55}$$

Therefore,

$$\mathbb{E}[Z_j^G] = G_S, \qquad \mathbb{E}[Z_j^\sigma] = \sigma_{S,\Pi}^2. \tag{56}$$

This construction is useful because reference noise does not enter the target moments. Cross-products between independent reference replicates remove the centered reference noise, leaving the directional signal and the candidate measurement error as the two quantities needed by SER.

Robust confidence bounds. The observations in Eqs. (52) and (53) contain products of directional readings and can have heavy tails. We therefore estimate their centers with a median-of-means estimator and use a Catoni-style upper bound on their variance.

Let $Z_1, \ldots, Z_J$ denote either the signal observations Z_j^G or the error observations Z_j^σ. We first form $n = J/2$ disjoint pairs and define

$$W_i = \frac{(Z_{a_i} - Z_{b_i})^2}{2}, \qquad i = 1, \ldots, n. \tag{57}$$

If $v = \mathrm{Var}(Z)$, then

$$\mathbb{E}[W_i] = v. \tag{58}$$

Assume that the Pearson kurtosis of Z is bounded by κ. Define

$$r_\kappa = \frac{\kappa + 1}{2}$$

and the monotone influence function

$$\psi(x) = \begin{cases} \log(1 + x + x^2/2), & x \geq 0, \\ -\log(1 - x + x^2/2), & x < 0. \end{cases} \tag{59}$$

For a one-sided scale error probability δ_s, let

$$a = \sqrt{\frac{2\log(1/\delta_s)}{n r_\kappa}}, \qquad B_C = \frac{n a^2 r_\kappa}{2} + \log\frac{1}{\delta_s}, \tag{60}$$

and

$$S(t) = \sum_{i=1}^{n} \psi\left(a\left(\frac{W_i}{t} - 1\right)\right). \tag{61}$$

The Catoni variance bound is

$$v_C = \inf\left\{t > 0 : S(t) \le -B_C\right\}. \tag{62}$$

Under the stated kurtosis condition,

$$\Pr(v \le v_C) \ge 1 - \delta_s. \tag{63}$$

We then partition the J observations into B disjoint blocks and compute the mean within each block. Let $\widetilde{\mu}$ be the median of these block means. If the smallest block contains $m_{\min}$ observations, choose $q < 1/2$ such that

$$\sum_{k=(B+1)/2}^{B} \binom{B}{k} q^k (1-q)^{B-k} \le \delta_c, \tag{64}$$

and define

$$r = \sqrt{\frac{v_C}{m_{\min} q}}. \tag{65}$$

Then, with probability at least $1 - \delta_s - \delta_c$,

$$|\widetilde{\mu} - \mathbb{E}[Z]| \le r. \tag{66}$$

Applying this procedure separately to the signal and error observations gives

$$\underline{G} = [\widetilde{\mu}_G - r_G]_+, \tag{67}$$

and

$$\overline{\sigma}^{\,2} = [\widetilde{\mu}_\sigma + r_\sigma]_+, \tag{68}$$

where $[x]_+ = \max(x, 0)$.

The lower bound is used for the signal and the upper bound for the error, making the resulting test conservative.

Certified SER ratio. We define the finite-sample certified SER ratio as

$$R_{\text{cert}} = \frac{\underline{G}}{\overline{\sigma}^{\,2}}. \tag{69}$$

The pair passes the finite-sample SER test when

$$R_{\text{cert}} > 1, \tag{70}$$

or equivalently,

$$\underline{G} > \overline{\sigma}^{\,2}. \tag{71}$$

On the joint confidence event,

$$G_S \geq \underline{G} > \overline{\sigma}^{\,2} \geq \sigma_{S,\Pi}^2. \tag{72}$$

Hence,

$$R_{\mathrm{cert}} > 1 \quad \Longrightarrow \quad R_{\mathrm{se}}(S, \Pi; \theta) > 1. \tag{73}$$

Theorem 1 then gives

$$\mathbb{E}[D\widehat{D}] > 0$$

for every admissible error under the corresponding mean-square bound. The finite-sample procedure therefore turns the population SER criterion into an operational test: a passing certificate shows that the directional signal remains detectable after accounting for both measurement error and finite-sample uncertainty.

In our experiments, we use $J = 1024$ directions and $K = 4$ reference replicates per direction. The center estimator uses eleven balanced median-of-means blocks, and the variance construction uses 512 disjoint pairs. We set the kurtosis bound to $\kappa = 15$. The eleven evaluated subspace–protocol pairs produce 22 signal/error statistics, and the confidence budgets are corrected jointly across this fixed family. The resulting simultaneous confidence level is 0.96.

The reference depends on the evaluated update space. Protocol T uses a forward-mode directional derivative under the fixed high-precision reference evaluator and applies to general candidate spaces. Protocol H is its closed-form specialization for an untied output head, computed from forward hidden states, probabilities, labels, and the sampled head direction. Under the same minibatch, normalization, and direction, both protocols measure the same directional quantity on the output-head subspace. These references are used only for SER analysis; standard QPZO training continues to use the two forward-loss evaluations described in Algorithm 1.

C More Results

C.1 Experimental Settings

Models and Datasets. We evaluate QPZO on three instruction-tuned language model families: Qwen2.5-7B-Instruct (Yang et al., 2025), Mistral-7B-Instruct-v0.3 (Jiang et al., 2023), and Gemma-2-9B-it (Team et al., 2024). For each model family, we use both GPTQ and AWQ 4-bit checkpoints, resulting in six model–quantizer settings. We load these quantized checkpoints directly and preserve their original low-bit backbone throughout QPZO training. This setting reflects the intended use of QPZO: starting from an already compressed model and directly adapting the continuously updateable parameters that are already available in the checkpoint.

We evaluate the models on five downstream NLP tasks covering both classification and generation. For classification, we use SST-2 (Socher et al., 2013; Wang et al., 2018) for sentiment analysis, RTE (Dagan et al., 2005; Wang et al., 2018) for textual entailment, WiC (Pilehvar & Camacho-Collados, 2019; Wang et al., 2019) for word-sense comparison, and BoolQ (Clark et al., 2019; Wang et al., 2019) for yes/no question answering. For generation, we use SQuAD (Rajpurkar et al., 2016) for extractive question answering. We report accuracy on the four classification tasks and F1 on SQuAD, with F1 used as the main SQuAD metric. Final task performance is measured on the complete validation split.

Baseline Methods. To evaluate the effectiveness of QPZO, we compare three practical ways to use the same 4-bit checkpoint. (1) **Frozen** directly evaluates the quantized checkpoint without training. (2) **QZO** (SHANG et al., 2026) performs zeroth-order fine-tuning through quantization scales. (3) **Full-dequantized MeZO** (Malladi et al., 2023) applies MeZO to the fully dequantized model. All trainable methods start from the same external 4-bit checkpoint, which keeps the initialization matched and lets the experiments isolate the effect of the update space itself.

Other Settings. We use 5,000 forward evaluations for classification tasks and 20,000 for SQuAD, with the same training budget and batch size applied to all ZO methods. And for each model family, we use the following specific checkpoints: Qwen2.5-7B-Instruct-GPTQ-Int4[1], Qwen2.5-7B-Instruct-AWQ[2], Mistral-7B-Instruct-v0.3-GPTQ[3], Mistral-7B-Instruct-v0.3-AWQ[4], Gemma-2-9B-it-GPTQ[5], and Gemma-2-9B-it-AWQ[6].

C.2 Stability of QPZO Across Random Seeds

Table 4: QPZO remains stable across random seeds

Model	Seed	SST-2	RTE	WiC	BoolQ	SQuAD
Qwen2.5-7B GPTQ	0	94.95	86.28	71.79	85.87	83.69
	1	94.61	87.36	71.00	86.36	82.85
	2	94.95	87.00	71.32	85.90	83.62
	Mean ± SD	94.84 ± 0.20	86.88 ± 0.55	71.37 ± 0.39	86.04 ± 0.27	83.39 ± 0.46
Qwen2.5-7B AWQ	0	95.07	89.89	69.59	83.91	84.82
	1	95.18	89.17	70.38	85.41	84.92
	2	95.18	89.53	70.53	84.77	84.86
	Mean ± SD	95.15 ± 0.07	89.53 ± 0.36	70.17 ± 0.50	84.70 ± 0.75	84.87 ± 0.05
Mistral-7B GPTQ	0	92.55	75.45	60.97	85.23	81.42
	1	94.15	75.09	57.99	84.53	80.89
	2	93.35	75.45	60.66	85.38	81.70
	Mean ± SD	93.35 ± 0.80	75.33 ± 0.21	59.87 ± 1.64	85.05 ± 0.46	81.34 ± 0.41
Mistral-7B AWQ	0	93.00	76.17	62.23	83.85	80.99
	1	93.69	77.26	58.78	83.82	81.23
	2	93.35	76.17	61.44	83.94	81.15
	Mean ± SD	93.35 ± 0.34	76.53 ± 0.63	60.82 ± 1.81	83.87 ± 0.06	81.12 ± 0.12
Gemma-2-9B GPTQ	0	95.30	84.48	65.52	83.20	90.94
	1	94.50	84.12	68.34	83.01	90.97
	2	94.61	84.84	66.93	82.80	90.99
	Mean ± SD	94.80 ± 0.43	84.48 ± 0.36	66.93 ± 1.41	83.00 ± 0.20	90.96 ± 0.03
Gemma-2-9B AWQ	0	95.18	83.39	68.18	81.70	90.34
	1	95.30	83.03	67.71	81.06	90.38
	2	95.53	83.75	68.50	81.36	90.39
	Mean ± SD	95.34 ± 0.18	83.39 ± 0.36	68.13 ± 0.39	81.37 ± 0.32	90.37 ± 0.03

Table 4 shows that QPZO remains stable across random seeds. We repeat QPZO with three random seeds under the same experimental settings. Across all datasets and model–quantizer settings, the standard deviation remains below 2 points, demonstrating strong stability and robustness to random-seed variation.

[1] https://huggingface.co/Qwen/Qwen2.5-7B-Instruct-GPTQ-Int4
[2] https://huggingface.co/Qwen/Qwen2.5-7B-Instruct-AWQ
[3] https://huggingface.co/thesven/Mistral-7B-Instruct-v0.3-GPTQ
[4] https://huggingface.co/thesven/Mistral-7B-Instruct-v0.3-awq
[5] https://huggingface.co/ModelCloud/gemma-2-9b-it-gptq-4bit
[6] https://huggingface.co/hugging-quants/gemma-2-9b-it-AWQ-INT4

C.3 Parameter Count Comparison Across Baselines

Table 5: Comparison of updated parameter counts across models and quantizers. MeZO updates the full model, QZO updates only quantization scales, and QPZO updates the complete QPS. Counts are reported in millions (M), with percentages relative to the full-model parameter count in parentheses.

Model	Quantizer	Full-model params	MeZO	QZO (scale-only)	QPZO (Full-QPS)
Qwen2.5-7B	GPTQ	7,615.617 M	7,615.617 M (100%)	50.979 M (0.67%)	1,090.199 M (14.32%)
	AWQ	7,615.617 M	7,615.617 M (100%)	50.979 M (0.67%)	1,090.199 M (14.32%)
Mistral-7B	GPTQ	7,248.024 M	7,248.024 M (100%)	54.526 M (0.75%)	268.702 M (3.71%)
	AWQ	7,248.024 M	7,248.024 M (100%)	54.526 M (0.75%)	268.702 M (3.71%)
Gemma-2-9B	GPTQ	9,241.706 M	9,241.706 M (100%)	65.028 M (0.70%)	918.110 M (9.93%)
	AWQ	9,241.706 M	9,241.706 M (100%)	65.028 M (0.70%)	918.110 M (9.93%)

Table 5 compares the number of parameters updated during training by the three trainable methods. QPZO updates substantially fewer parameters than full-dequantized MeZO, while maintaining highly competitive downstream performance. QZO uses only 0.67%–0.75% of the full-model parameters, but this extremely small update space is accompanied by considerable performance degradation and unstable adaptation across settings. QPZO instead uses a moderately larger checkpoint-native space, leading to substantially stronger and more consistent performance than QZO, while increasing peak training memory by only a quite small margin.

C.4 Pathwise Evolution of QPS Detectability

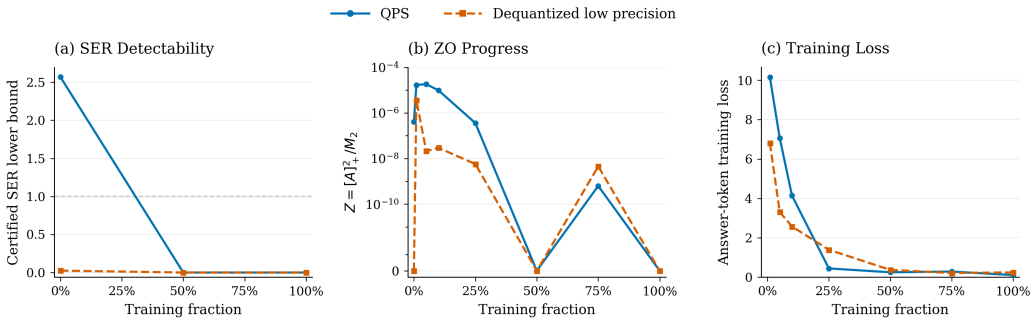


Figure 3: Pathwise evolution of QPS detectability on Qwen2.5-7B GPTQ/SST-2. Panel (a) reports the certified SER lower bound at the evaluated checkpoints, with the dashed line marking the $R_{\text{se}} = 1$ threshold. Panel (b) tracks the progress score $Z = [A]_+^2/M_2$, and Panel (c) reports the answer-token training loss begin at the first checkpoint.

Figure 3 follows the update spaces throughout training to examine how detectability evolves with the model state. At initialization, QPS has a certified SER lower bound well above one, while the dequantized low-precision space remains close to zero. QPS also maintains a substantially larger Z during the early training stage, showing that its measurable directional signal translates into stronger usable ZO progress.

The SER behavior is also reflected in the loss trajectory. QPS starts with a clear SER margin and rapidly reduces the training loss during the early stage of adaptation, reaching a low-loss region

within the first quarter of training. As optimization proceeds, both the SER lower bound and Z decrease after most of the loss reduction has already occurred. This pattern shows that SER tracks the local availability of useful forward-query signal: strong detectability appears when substantial optimization progress remains, while the measurable directional signal becomes weaker as the model approaches a low-loss region. Together, the three panels connect QPS detectability with effective ZO progress and the resulting loss reduction.

C.5 Generalization of QPS Detectability

Table 6: SER evaluation of QPS detectability across tasks, models, and quantizers.

Model	Quantizer	Task	Directions	Certified SER	Result
Panel (a): Cross-task evaluation					
Qwen2.5-7B	GPTQ	SST-2	1024	3.021	✓
		RTE	1024	3.486	✓
		WiC	1024	4.845	✓
		BoolQ	1024	1.799	✓
		SQuAD	1024	1.344	✓
Panel (b): Cross-quantizer and cross-model evaluation on SST-2					
Qwen2.5-7B	GPTQ	SST-2	1024	3.021	✓
Qwen2.5-7B	AWQ	SST-2	1024	5.511	✓
Mistral-7B	GPTQ	SST-2	1024	2.397	✓

The SER experiments in the main text focus primarily on a single setting. We next examine whether the detectable forward-query signal in QPS extends beyond this setting. We evaluate five tasks on Qwen-GPTQ and additionally evaluate SST-2 on Qwen-AWQ and Mistral-GPTQ. Table 6 shows that the certified SER exceeds one in every evaluated setting, indicating that QPS consistently exhibits detectable task-relevant directional signals across different tasks, quantizers, and model families.