AutoQRA

Joint Optimization of Mixed-Precision Quantization and Low-rank Adapters

Changhai Zhou, Shiyang Zhang, Yuhua Zhou, Qian Qiao, Jun Gao, Cheng Jin, Kaizhou Qin, Weizhong Zhang
Fudan University | Yale University | Zhejiang University

+1.15 to +3.44
avg score gain over QLoRA

12 to 22%
lower measured memory

4 / 4
best <=4-bit backbones

18x
fewer HF evals

One memory budget is shared by quantized backbone weights, LoRA adapters, optimizer state, scales, and zero-points.

github.com/harrysyz99/autoqra

Why joint allocation matters

One budget, two coupled knobs

Coupled knobs

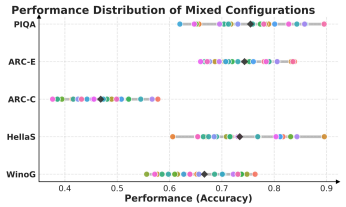
Lower precision creates noise; LoRA rank can repair part of it during learning.

Proxy mismatch

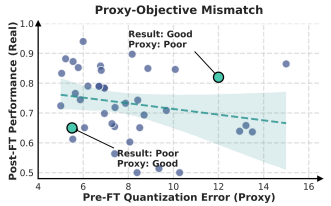
Frozen reconstruction quality does not reliably predict post-tuning utility.

Black-box search

Useful scores require partial fine-tuning, so exhaustive search is too expensive.

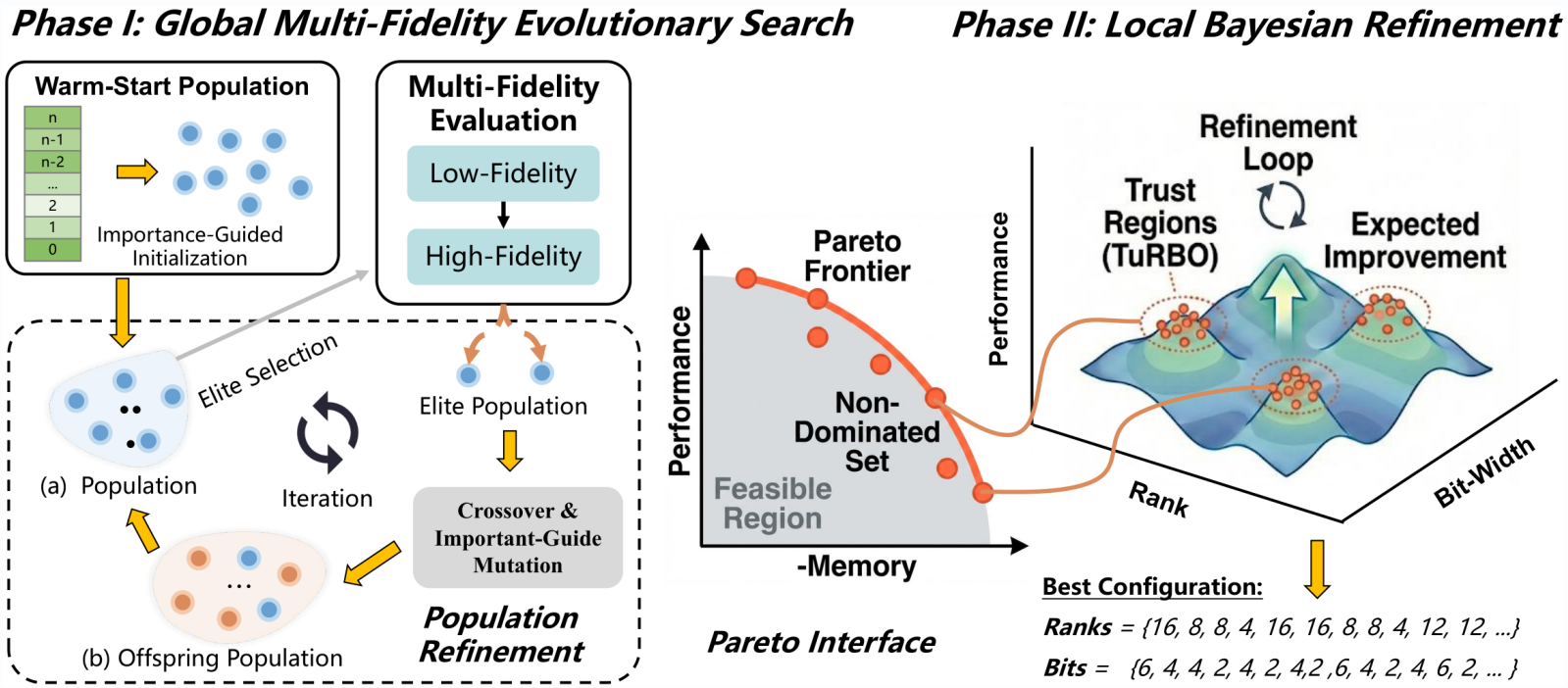


Same memory, different utility: mixed bit-rank configurations diverge after tuning.



Frozen proxy rankings only moderately agree with final scores: rho is about 0.46.

AutoQRA: measured coarse-to-fine search



Warm start
sample feasible bit-rank configs

Global EA
screen broad candidate space

Local BO
refine high-utility regions

Select
return deployable model

Phase I builds a repair-aware frontier cheaply; Phase II spends high-fidelity tuning near promising neighborhoods.

Main result: more accuracy per GB

+1.15 to +3.44
higher average score

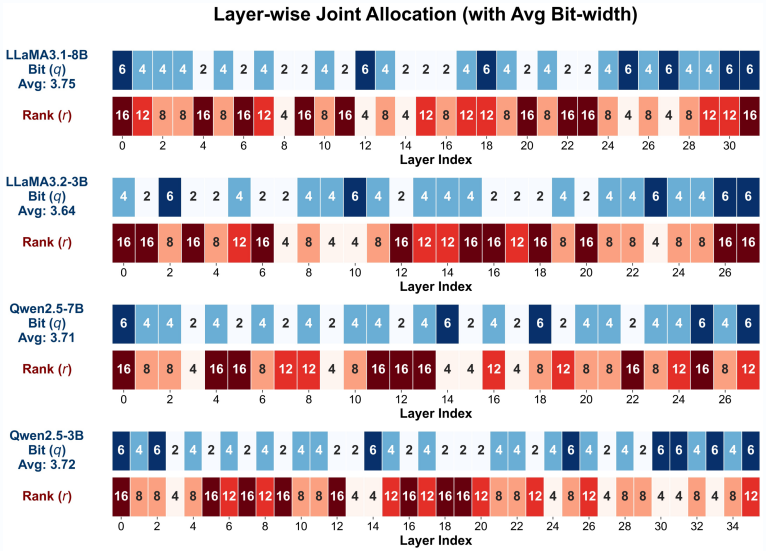
12 to 22%
lower memory footprint

3.64 to 3.75
average bit-width

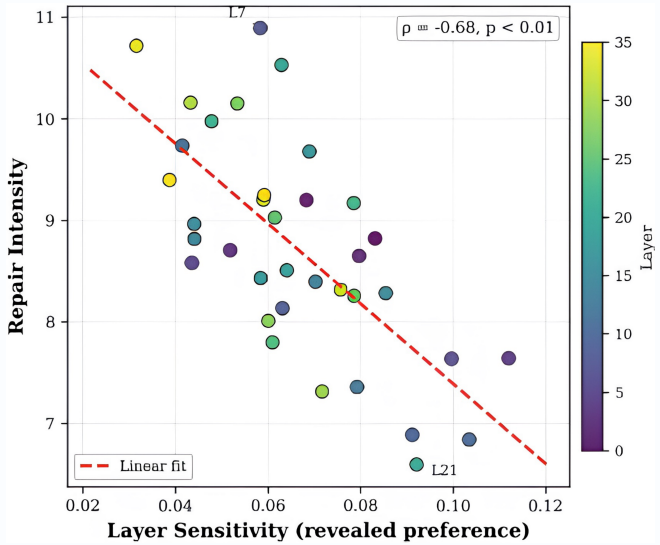
Backbone	QLoRA Avg	AutoQRA Avg	Delta Avg	QLoRA GB	AutoQRA GB	Delta GB
LLaMA3.1-8B	67.45	**69.83**	**+2.38**	15.22	**13.78**	**-1.44**
LLaMA3.2-3B	64.43	**65.58**	**+1.15**	8.90	**6.72**	**-2.18**
Qwen2.5-7B	69.01	**71.35**	**+2.34**	15.60	**13.05**	**-2.55**
Qwen2.5-3B	62.89	**66.33**	**+3.44**	8.16	**6.75**	**-1.41**

AutoQRA <=4-bit improves task-average accuracy and measured memory on every tested backbone. The opt-budget variant further recovers or exceeds FP16 LoRA quality.

What the learned allocations reveal



Lower-bit layers often receive higher LoRA ranks after joint search.



Repair edits hit low-sensitivity layers, removing less damaging capacity.

Rank repairs low bits

The search trades redundant precision for adapter capacity where learning can help.

Structured repair

Layer repair frequency has rho = -0.68 with sensitivity, avoiding brittle layers.

Tighter budgets

0.7x budget keeps Avg 63.92 vs 64.55 and HumanEval 41.39 vs 41.96.

Takeaways for the poster session

1. Joint budget

Bit-width and LoRA rank are coupled by memory and learning dynamics; fixing one first can misallocate capacity.

2. Measured search

Static reconstruction proxies miss post-tuning utility, so AutoQRA uses multi-fidelity measured evaluations.

3. Useful structure

Discovered configs avoid brittle task drops and reveal compensation between precision and adapter rank.

Check	Comparison	Score
HC3 reuse	Alpaca transfer / QLoRA	**65.54 / 62.45**
Task rescue	WinoG QLoRA / AutoQRA	**63.30 / 73.30**
HumanEval	AutoQRA Opt / FP16 LoRA	**43.07 / 43.19**
14B low-bit	AutoQRA <=4-bit / QLoRA	**80.30 / 79.69**
14B opt	AutoQRA Opt / FP16 LoRA	**80.71 / 80.31**