

TRAINING DYNAMICS IMPACT POST-TRAINING QUANTIZATION ROBUSTNESS

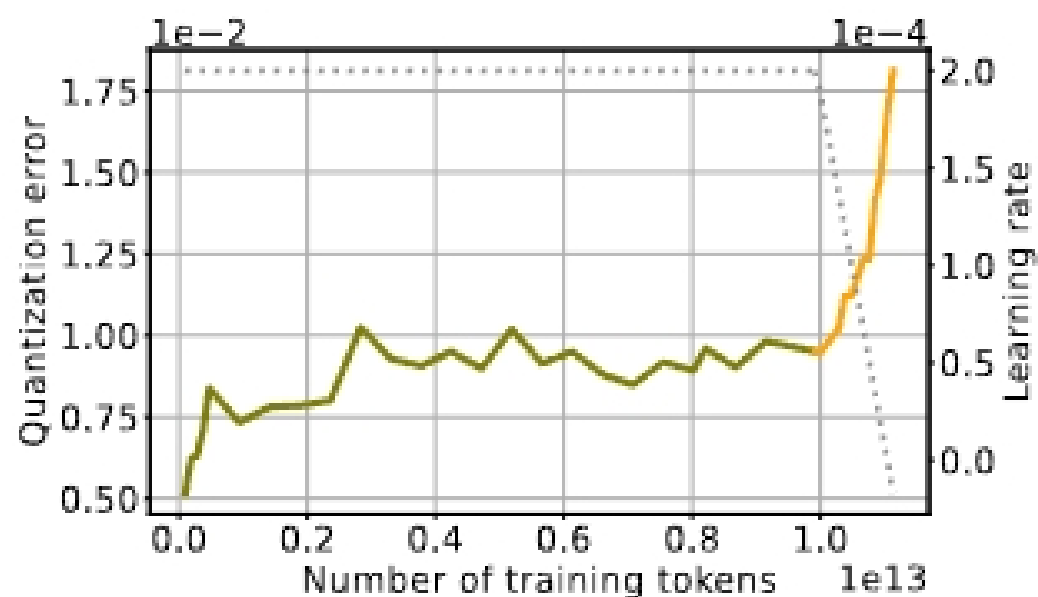
1.研究背景

大模型量化的鲁棒性难题

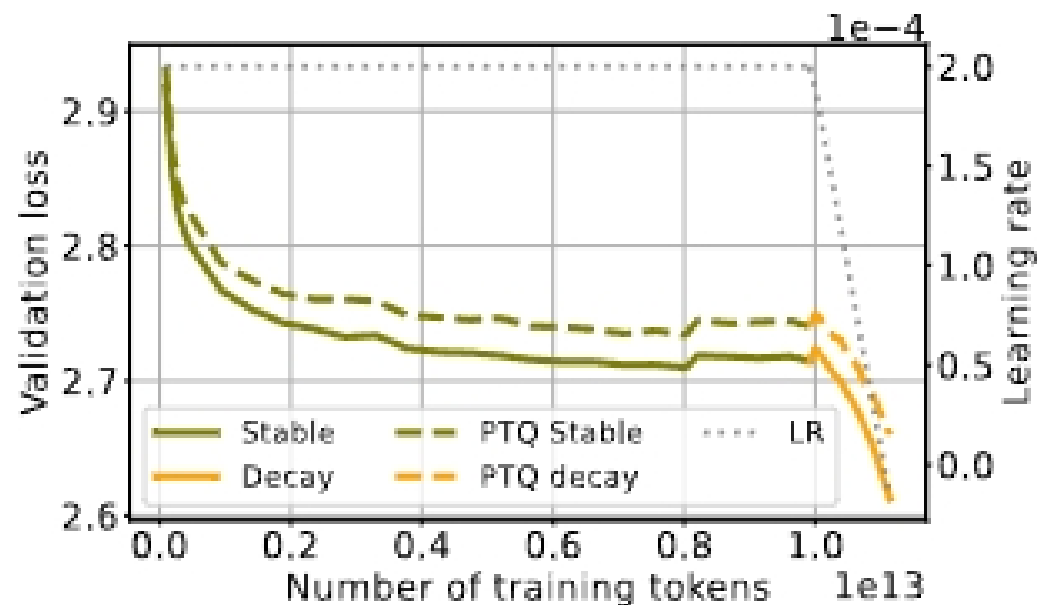
早期研究提出：训练数据量的增加会加剧模型的量化退化

本文认为：上述现象并非由数据量直接导致，而是训练过程中被忽略的超参数所造成的假象

2.核心发现



(a) 4-bit quantization error vs training tokens.



(b) Validation loss vs training tokens.

SmolLM3训练后各阶段的3bit量化效果

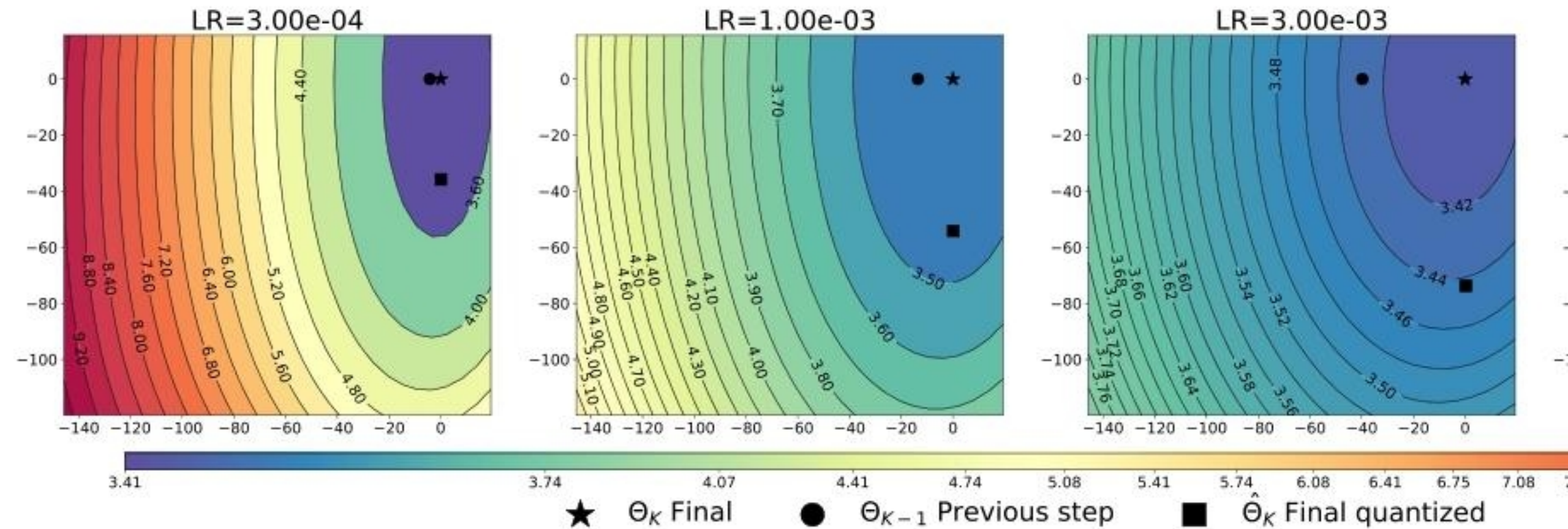
3.底层机制

损失景观 (loss landscape)：将模型的所有可能权重视为一个多维空间，空间中每个点的高度就是这个权重对应的损失值 (loss)

目标要在这个空间中找到高度最低的“山谷”

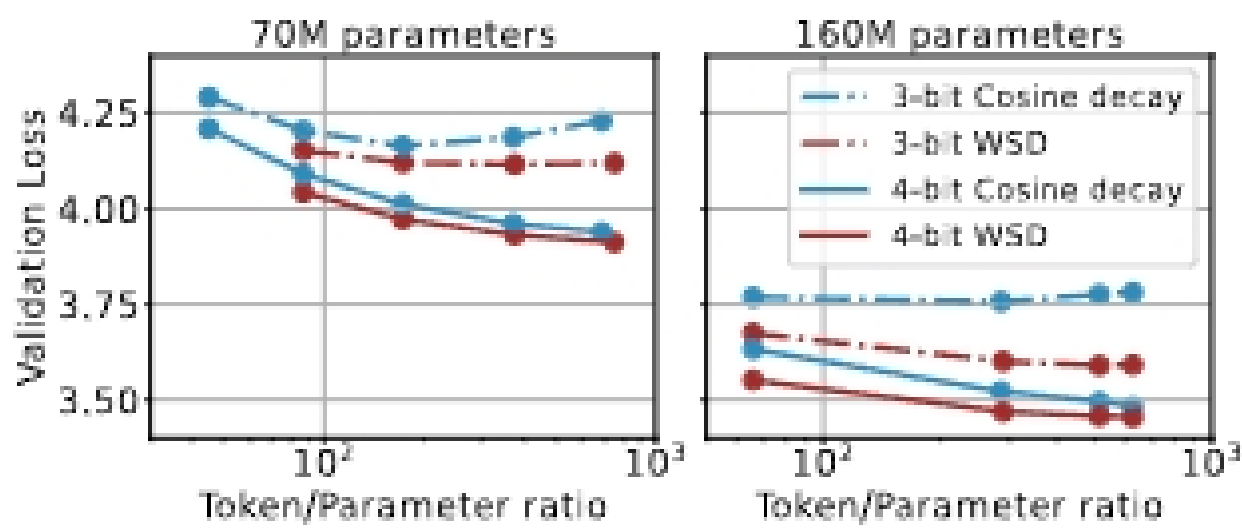
量化鲁棒性的本质：量化会给权重带来微小扰动。若“山谷”两侧较为平缓则扰动带来的影响较小，既鲁棒性高

3.底层机制

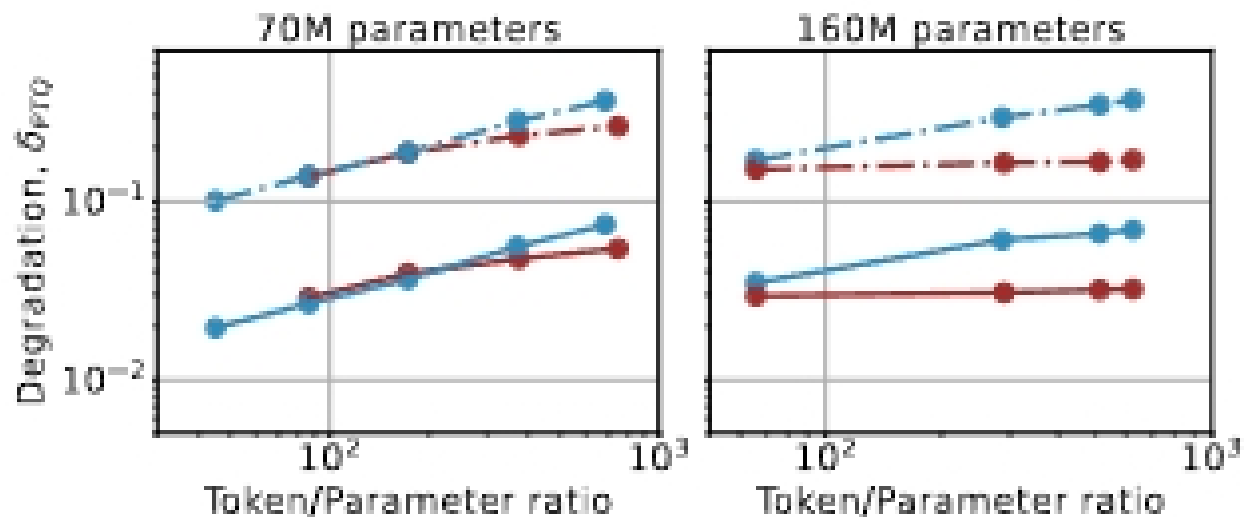


4. 优化策略

调整学习率调度策略、提高峰值学习率



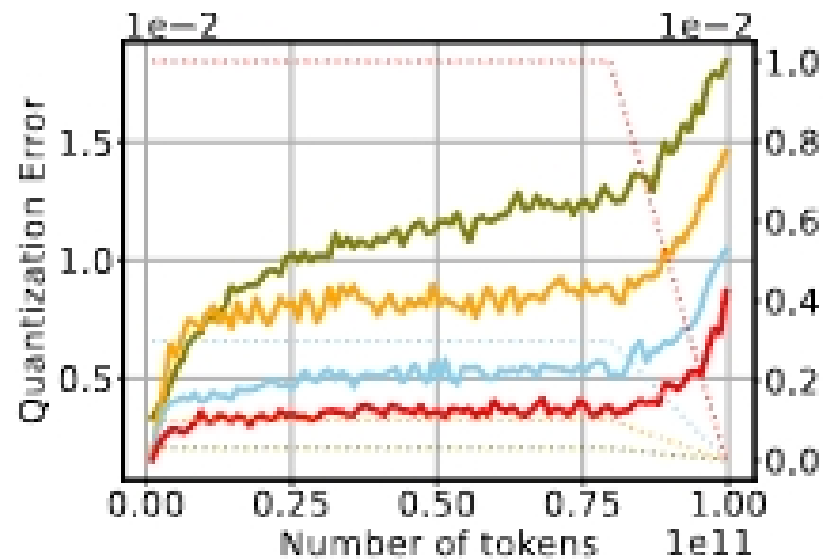
(a) Validation loss.



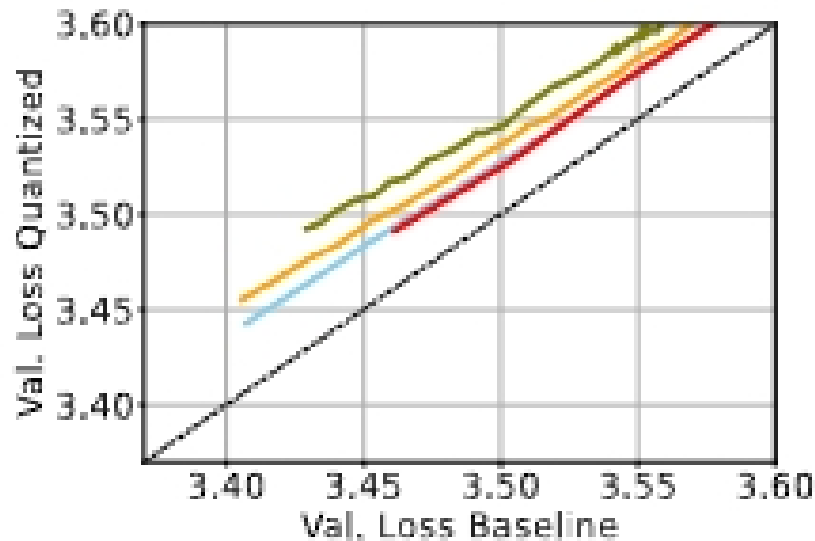
(b) $\delta_{PTQ} := \text{CE}(\hat{W}) - \text{CE}(W)$.

4. 优化策略

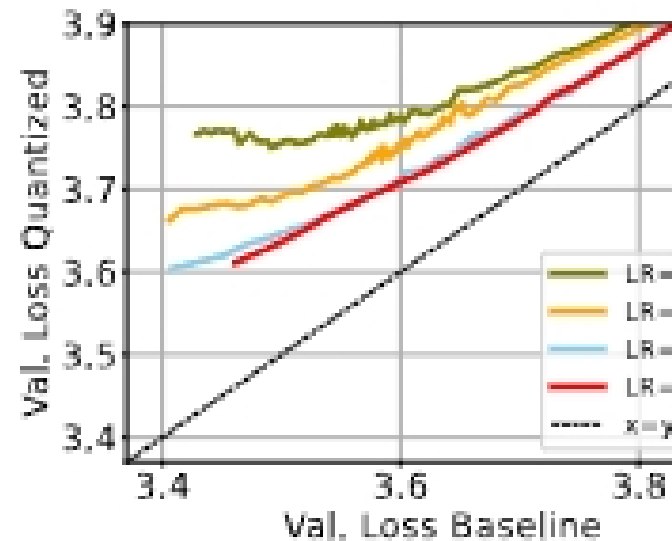
调整学习率调度策略、提高峰值学习率



(a) 4-bit quantization error.



(b) FP to 4-bit validation loss.



(c) FP to 3-bit validation loss.

5.权重平均

LAWA（单轨迹权重平均）：对WSD训练稳定阶段保存多个连续的Checkpoint进行权重的算数平均

Model Soup（多轨迹权重融合）：融合不同训练轨迹得到的多个模型权重