Intriguing Properties of Quantization at Scale

Arash Ahmadian*†§, Saurabh Dash*‡, Charlie Chen*‡, Bharat Venkitesh†, Stephen Gou†, Phil Blunsom†, Ahmet Üstün‡, Sara Hooker‡

* Equal Contribution , † Cohere For AI, ‡ Cohere, § University of Toronto

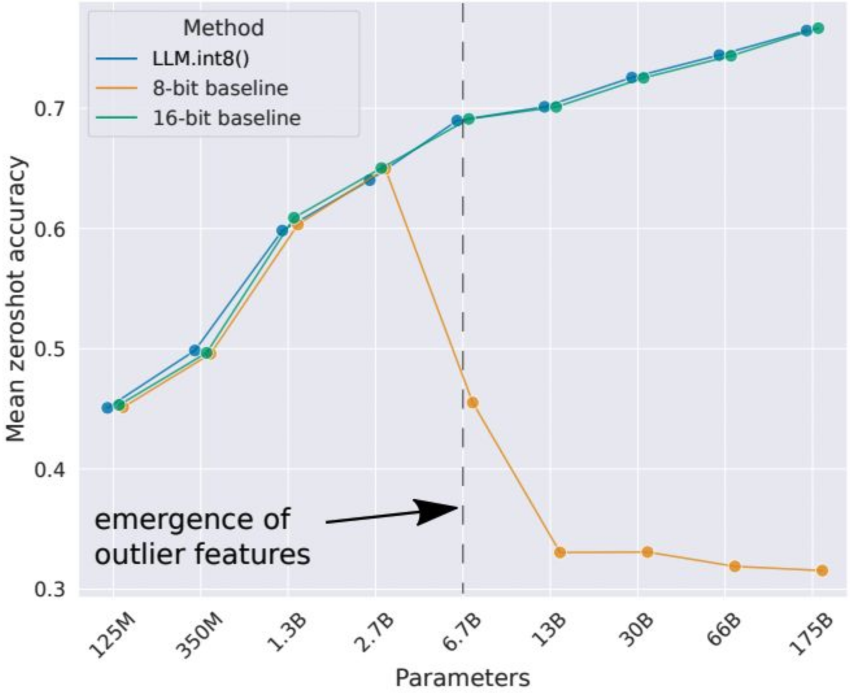
Email: arash@cohere.com

Cohere For AI

Traditional INT8 Methods FAIL at Scale

- **Emergent outlier dimensions** in LLMs' hidden-states make **Post Training Quantization (PTQ)** difficult for models at scale (> 6B).

- LLM.int8() → fixes performance drop **but** is not <u>easily generalizable</u> & no <u>latency benefit</u>



Credits: Dettmers et al., 2022

“

Are emerging properties of LLM truly inherent to scale, or can they be altered and conditioned by optimization choices?

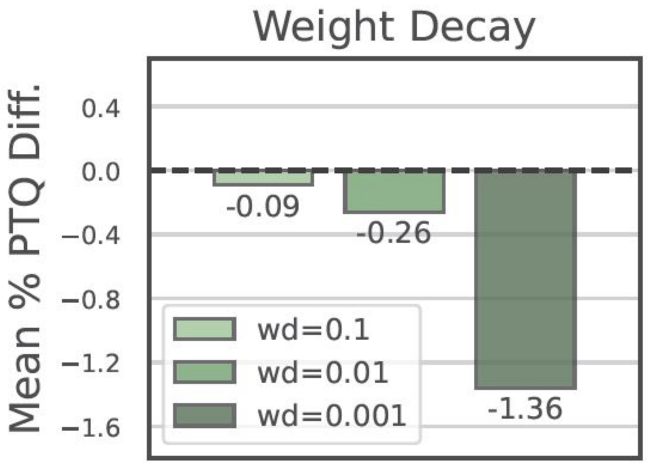
Nurture: Optimization Choices

- Isolate effects of each pre-training optimization choices:
  - Control other choices while varying one
  - Due to high cost of training at scale → 6B early checkpoint (75k steps)
  - Quantize **both hidden-states and weights** → measure degradation

Experimental Axes	Choices
Weight decay	0.001, 0.01, 0.1
Gradient clipping	None, 1
Dropout	0, 0.1, 0.4, 0.8
Half-precision	bf16, fp16

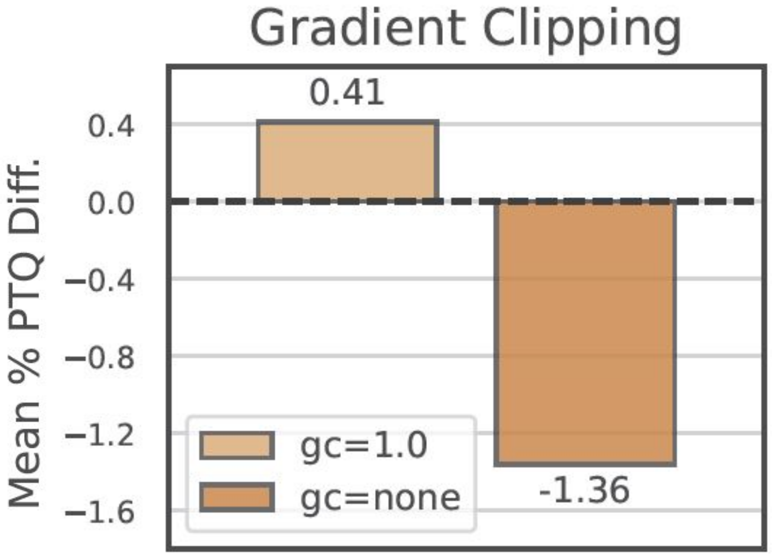
Weight Decay

- Vary weight decay with gradient-clipping turned off
- Want to decouple their effects
- Higher weight decay → better PTQ



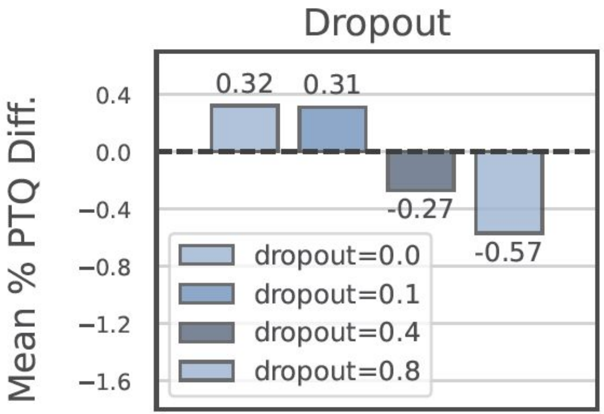
Gradient Clipping

- Vary gradient-clipping with weight decay = 0.001
- Want to decouple the effects of two
- Gradient Clipping → better PTQ



Dropout

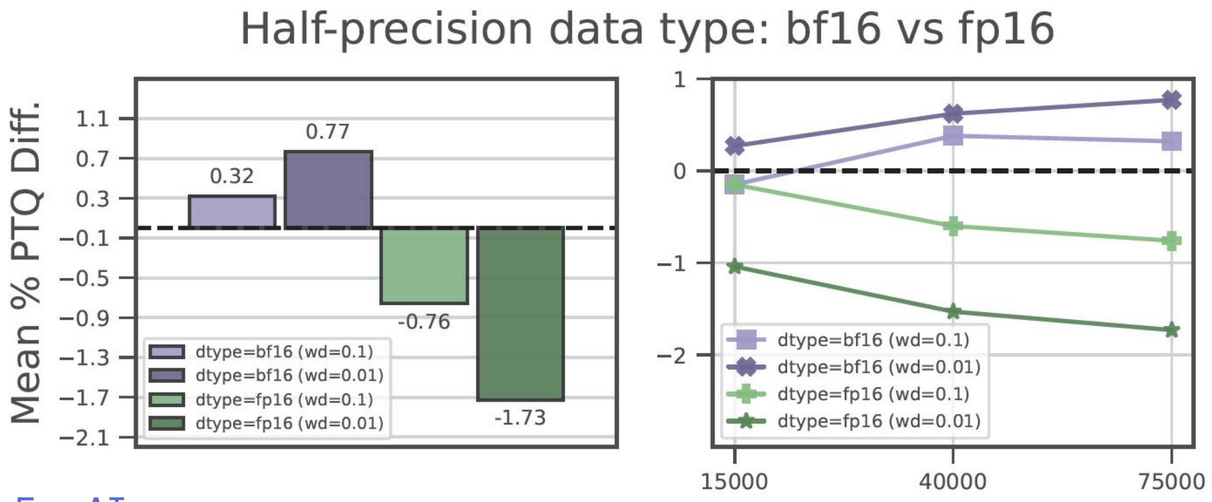
- Only applied to the hidden-states right before a residual connection
- Not applied to embeddings
- Smaller dropout → better PTQ
- dropout=0.8 has significantly worse performance before quantization(expected)



BF16 > FP16

- FP16 → worse PTQ (most significant out of all experimental axis)

- Degradation trends are consistent over time



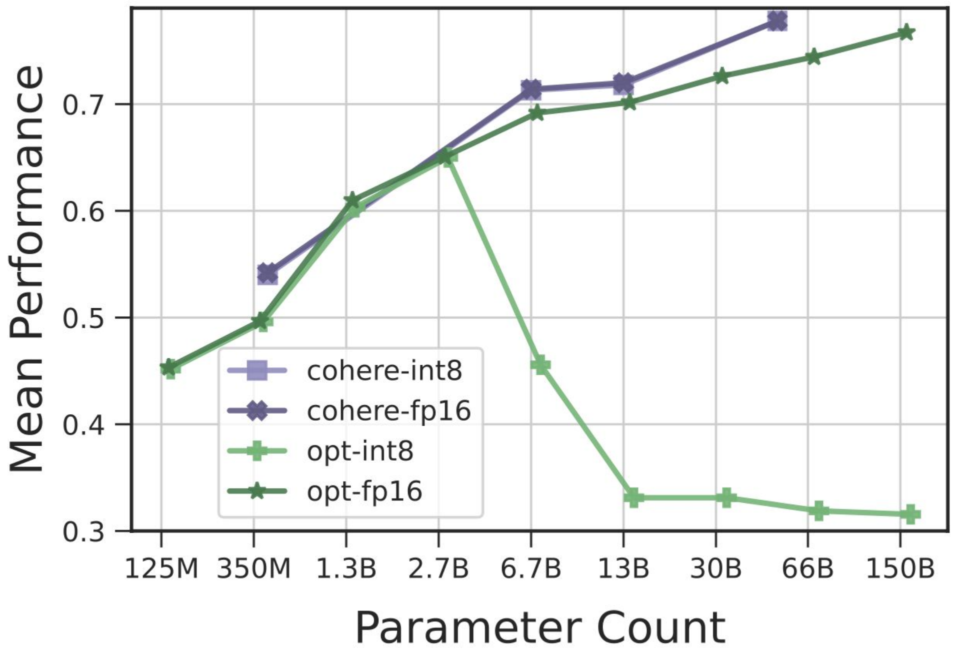
Cohere For AI

Validation at Scale

- Q: Do insights at 6B translate to other scales?
- A: **Yes!**

Model	Degradation
cohere-52B	**0.0%**
OPT-66B	42%*

Weight decay	Gradient clipping	Dropout	Half-precision datatype
0.1	1.0	0	bf16



* directly taken from Dettmers et al., 2022

Final Takeaways

- Outliers at scale are due to **nurture** rather than **nature**
- Train with **bf16**, gradient clipping, higher weight decay, and low dropout
- Simple INT8 quantization of both hidden-states and weights is feasible at scale

Email: arash@cohere.com

Checkout the paper!

Cohere For AI