Why Do Some Inputs Break Low-Bit LLM Quantization?

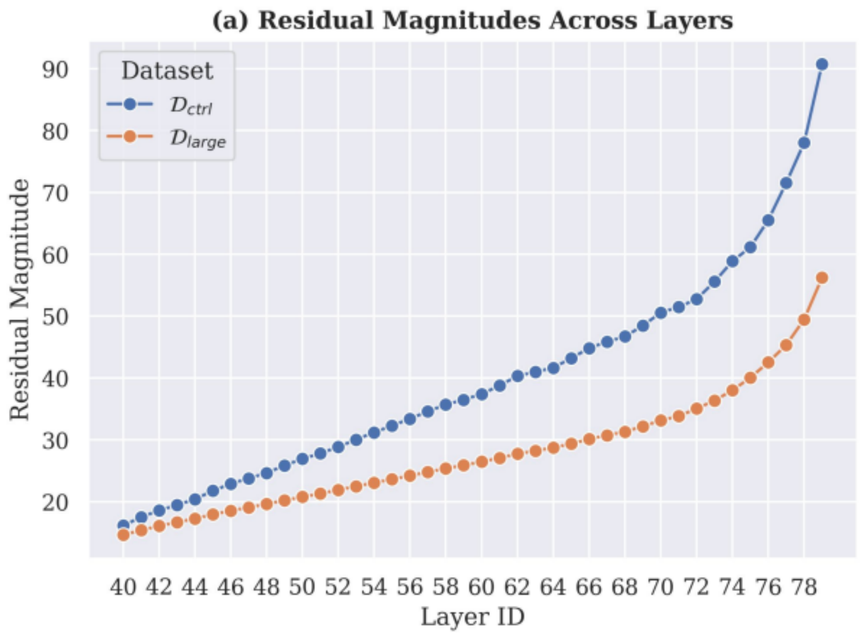
Ting-Yun Chang, Muru Zhang Jesse Thomason, Robin Jia

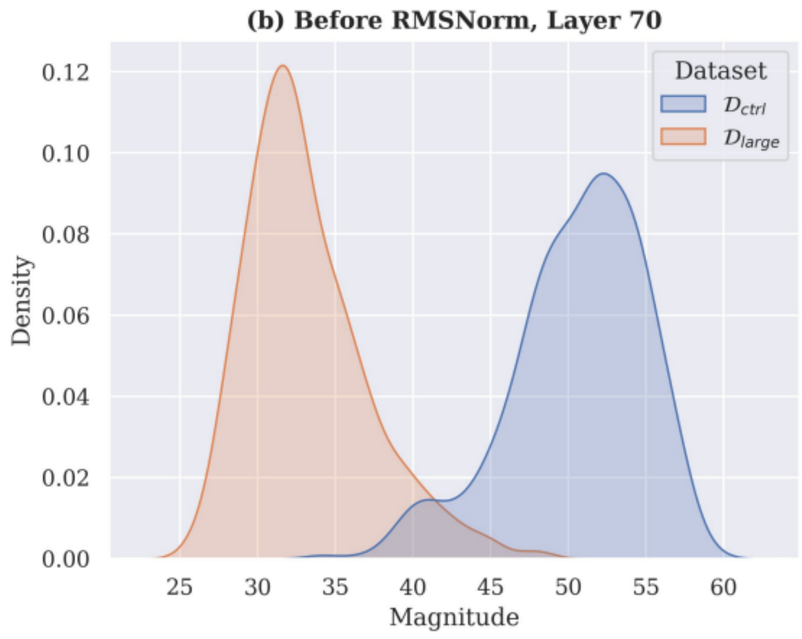
USC Viterbi University of Southern California

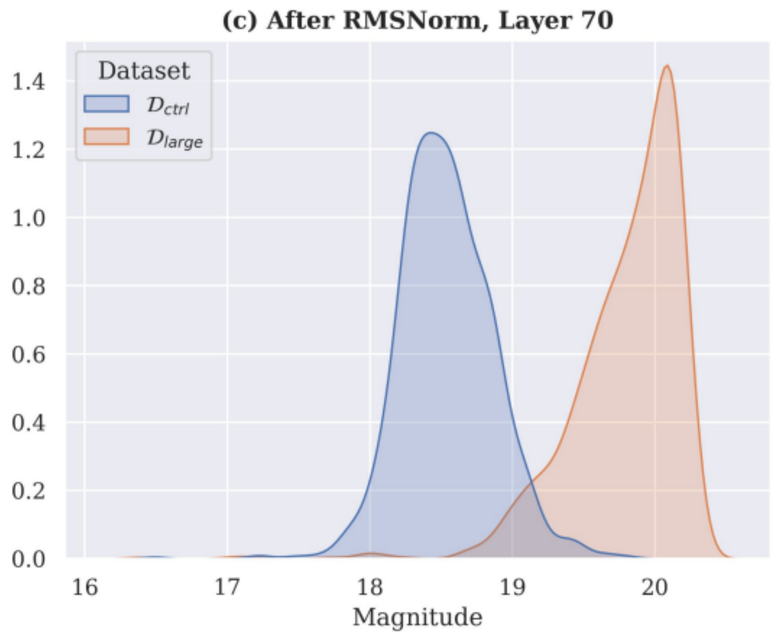
- LLM quantization is awesome! Prior work mainly focuses on method development.
- We conduct a thorough analysis on **understanding errors in 3-4 bit weight-only quantization**
  - Where do the errors stem from? What kinds of examples tend to have larger degradation?

Quantization Disproportionately Affect Certain Examples

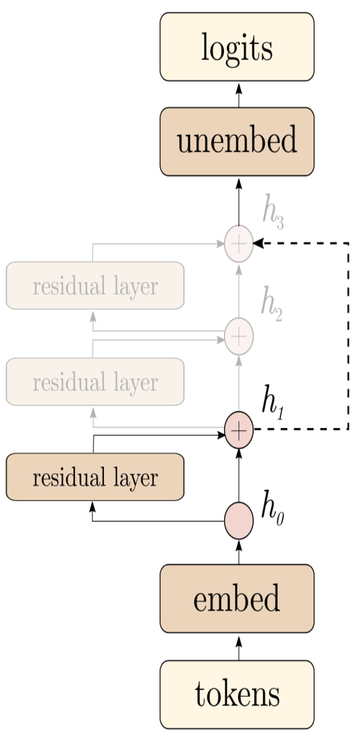
- **The quantization errors of various pairs of methods are highly correlated (avg. ρ = 0.82)**
  - Methods: **AWQ**, **NormalFloat** (from QLoRA), **GPTQ**, **EfficientQAT**
  - Quantization errors: ΔNLL and KL(BF16, Quantized) on FineWeb sequences
  - Model: Qwen2.5-7B, Llama3-8B, Mistral-Nemo-12B, Llama3-70B
- **Certain examples yield large errors across diverse methods**
- **Residual magnitudes from the full-precision model is predictive of quantization errors**

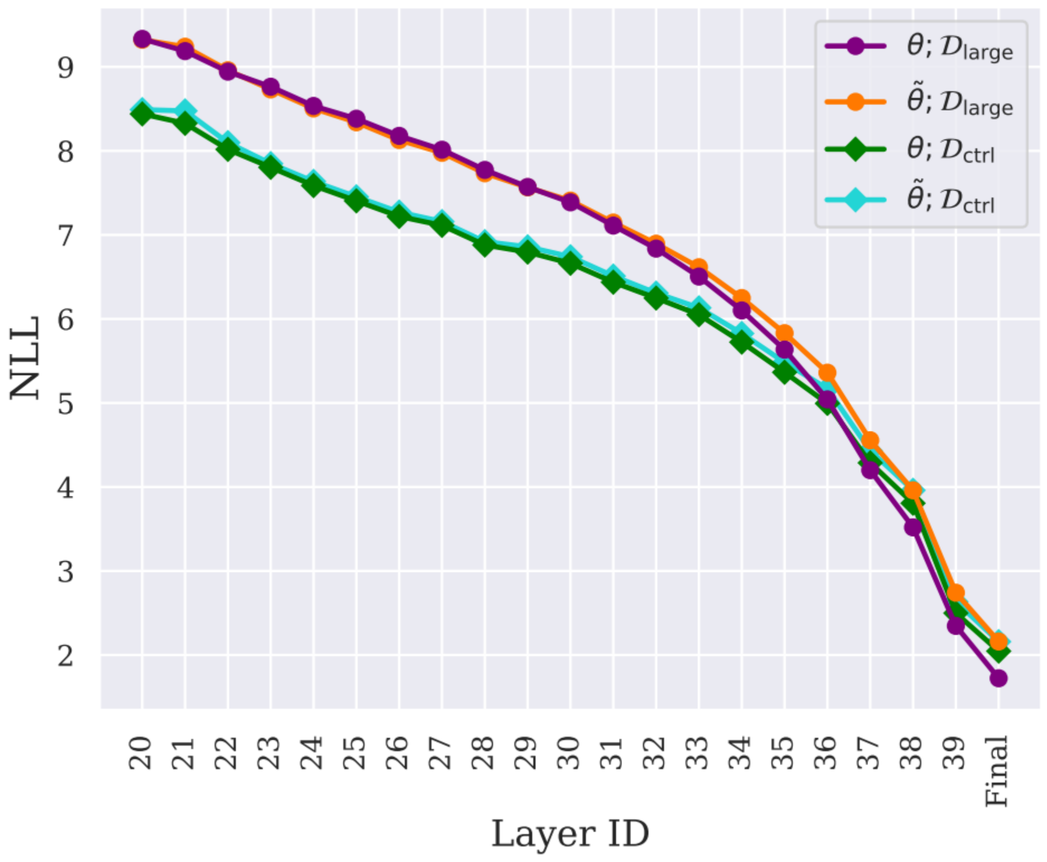


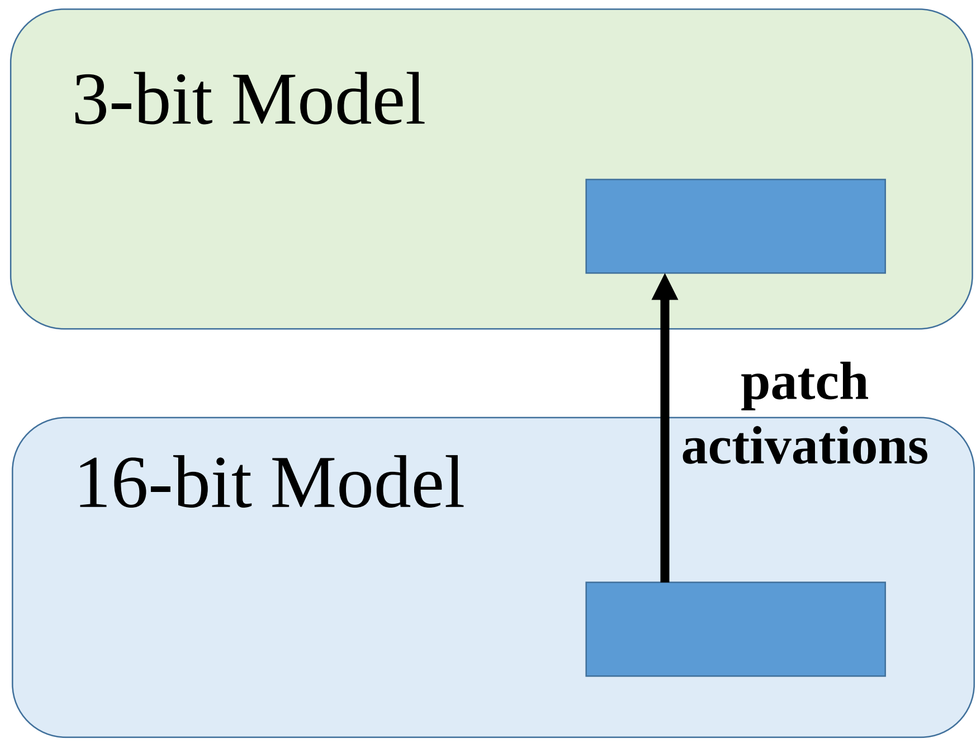




Which Parts of the Model Lead to Large Errors? Early Exiting, Cross-Model Patching, Weight Recovering







Large-error examples rely on precise **activations**, not weights, of the **upper layers** and **MLP-gate** outputs

What Kinds of Data Suffer From Large Quantization Errors?