LOFTQ (LoRA-Fine-Tuning-aware Quantization) is a novel quantization framework designed to enhance the performance of pre-trained language models when both quantization and LoRA fine-tuning are applied. This framework addresses the performance gap observed between full fine-tuning and the combination of quantization with LoRA fine-tuning by providing a better initialization for LoRA. LOFTQ integrates low-rank approximation with quantization to improve alignment with the original high-precision weights, thus enhancing generalization in downstream tasks. It is particularly effective in low-bit scenarios, such as 2-bit and 2/4-bit mixed precision regimes, and consistently outperforms existing methods like QLoRA. The framework has been evaluated across various tasks, including natural language understanding, question answering, summarization, and natural language generation, demonstrating significant improvements in performance.

QLoRA Performance with Different Bits

Number of Bits	Log of Perplexity (a)	Log of Perplexity (b)
16	2.44	1.0
8	2.01	1.0
4	2.23	1.0
3	2.83	1.0
2.5	11.37	2.82
2.25	11.48	6.40
2	11.30	7.10