LoftQ, or LoRA-Fine-Tuning-aware Quantization, is a novel quantization framework designed for pre-trained language models that require both quantization and LoRA fine-tuning. It addresses the performance gap observed when quantization and LoRA fine-tuning are applied together, compared to full fine-tuning. LoftQ integrates low-rank approximation with quantization to better align with the original high-precision pre-trained weights, providing a more effective initialization for LoRA fine-tuning. This approach significantly improves generalization in downstream tasks such as natural language understanding, question answering, summarization, and natural language generation.

LoftQ has been shown to outperform existing quantization methods, particularly in challenging low-bit scenarios like 2-bit and 2/4-bit mixed precision regimes. It achieves notable improvements in performance metrics, such as a 1.1 and 0.8 gain in Rouge-1 for XSum and CNN/DailyMail, respectively, and over an 8% gain on MNLI and more than 10% on SQuADv1.1 with both 2-bit NormalFloat and 2-bit uniform quantization. The framework effectively mitigates the initialization discrepancy introduced by quantization, especially at lower bit levels, ensuring better task adaptation and 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