LOFTQ (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. Experiments demonstrate that LOFTQ outperforms existing quantization methods, particularly in low-bit scenarios like 2-bit and 2/4-bit mixed precision regimes. The framework is shown to be effective across various quantization methods, achieving notable performance gains on tasks like MNLI and SQuADv1.1.