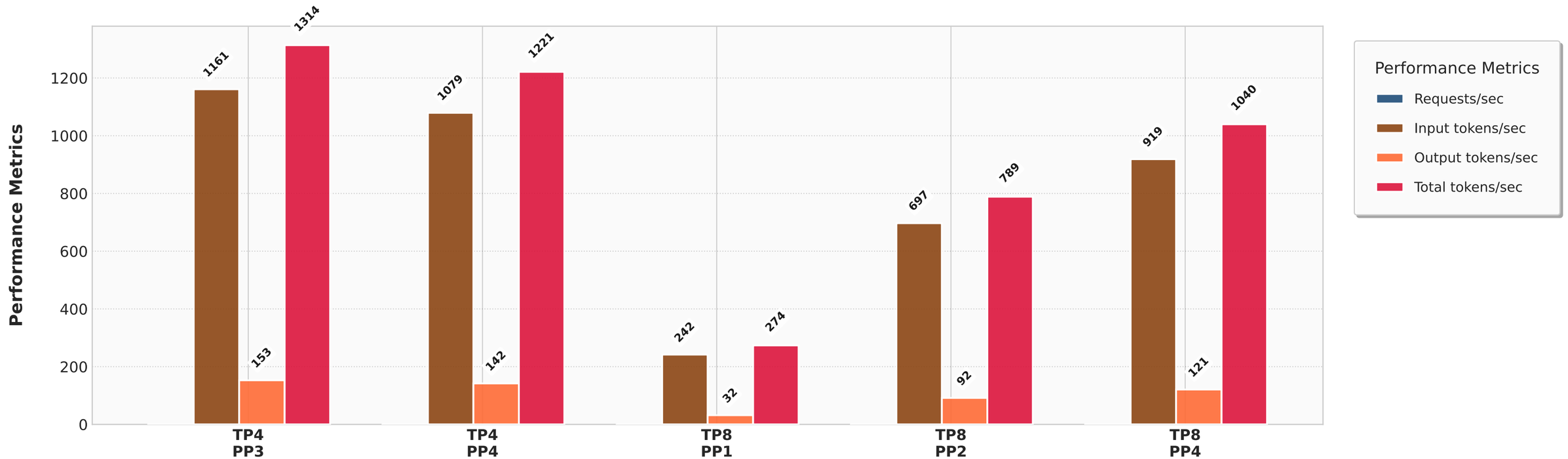
DeepSeek-R1-Distill-Llama-70B Performance Analysis
TP/PP Parallelism Configurations (Input: 8192 tokens, Output: 2048 tokens)



Parallelization Scaling Efficiency & Cost Efficiency Analysis