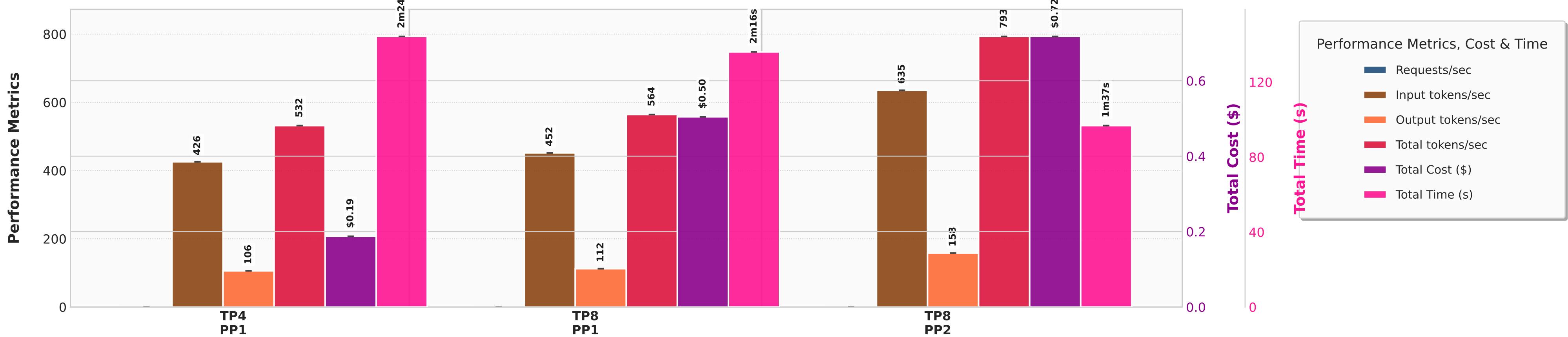
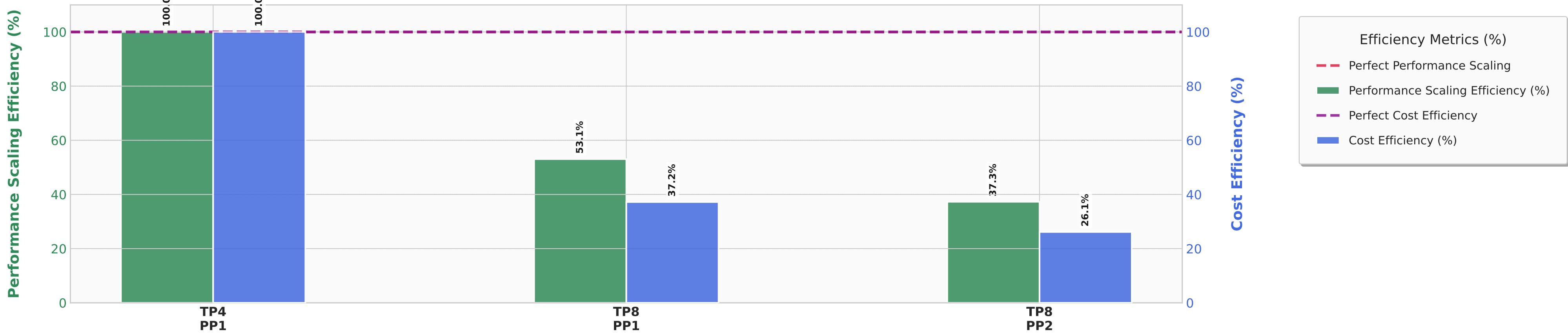


# DeepSeek-R1-Distill-Llama-70B Performance Analysis

## TP/PP Parallelism Configurations (Input: 2048 tokens, Output: 512 tokens)



## Performance Scaling Efficiency & Cost Efficiency Analysis



## GPU Utilization Analysis

