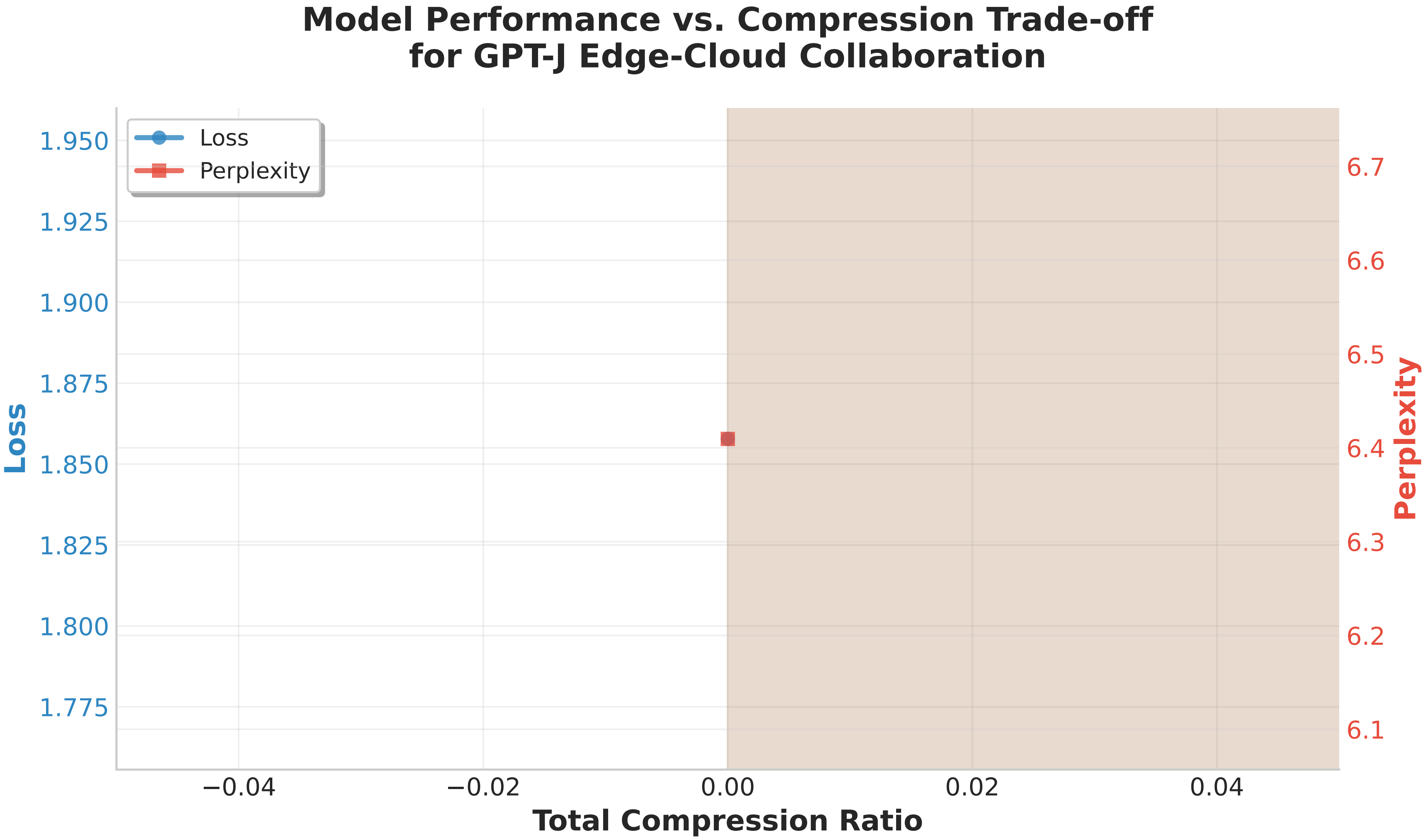
Model Performance vs. Compression Trade-off
for GPT-J Edge-Cloud Collaboration
Loss
Perplexity
Loss
1.950
1.925
1.900
1.875
1.850
1.825
1.800
1.775
Perplexity
6.7
6.6
6.5
6.4
6.3
6.2
6.1
−0.04
−0.02
0.00
0.02
0.04
Total Compression Ratio