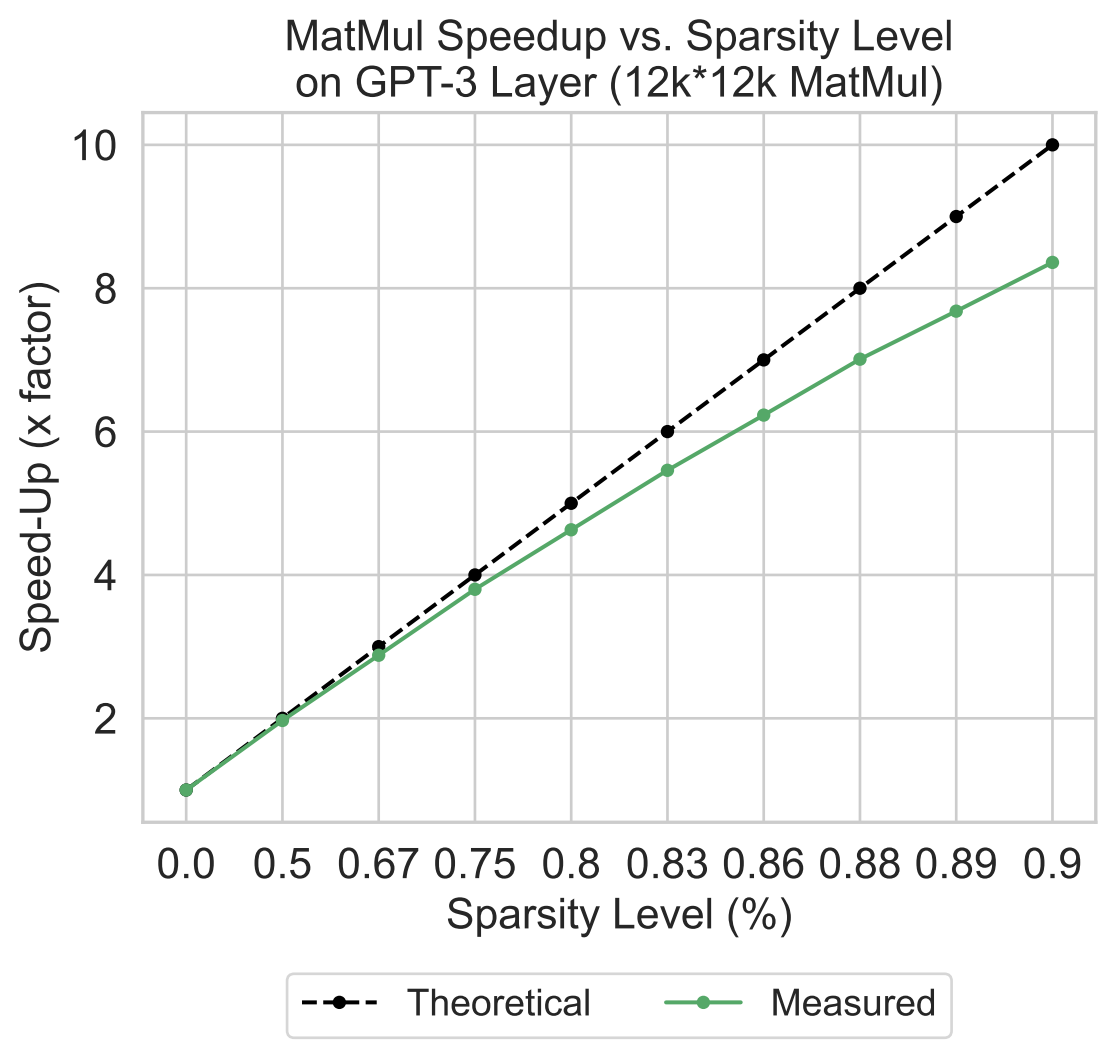
MatMul Speedup vs. Sparsity Level
on GPT-3 Layer (12k*12k MatMul)
Speed-Up (x factor)
10
8
6
4
2
0.0
0.5
0.67
0.75
0.8
0.83
0.86
0.88
0.89
0.9
Sparsity Level (%)
Theoretical
Measured