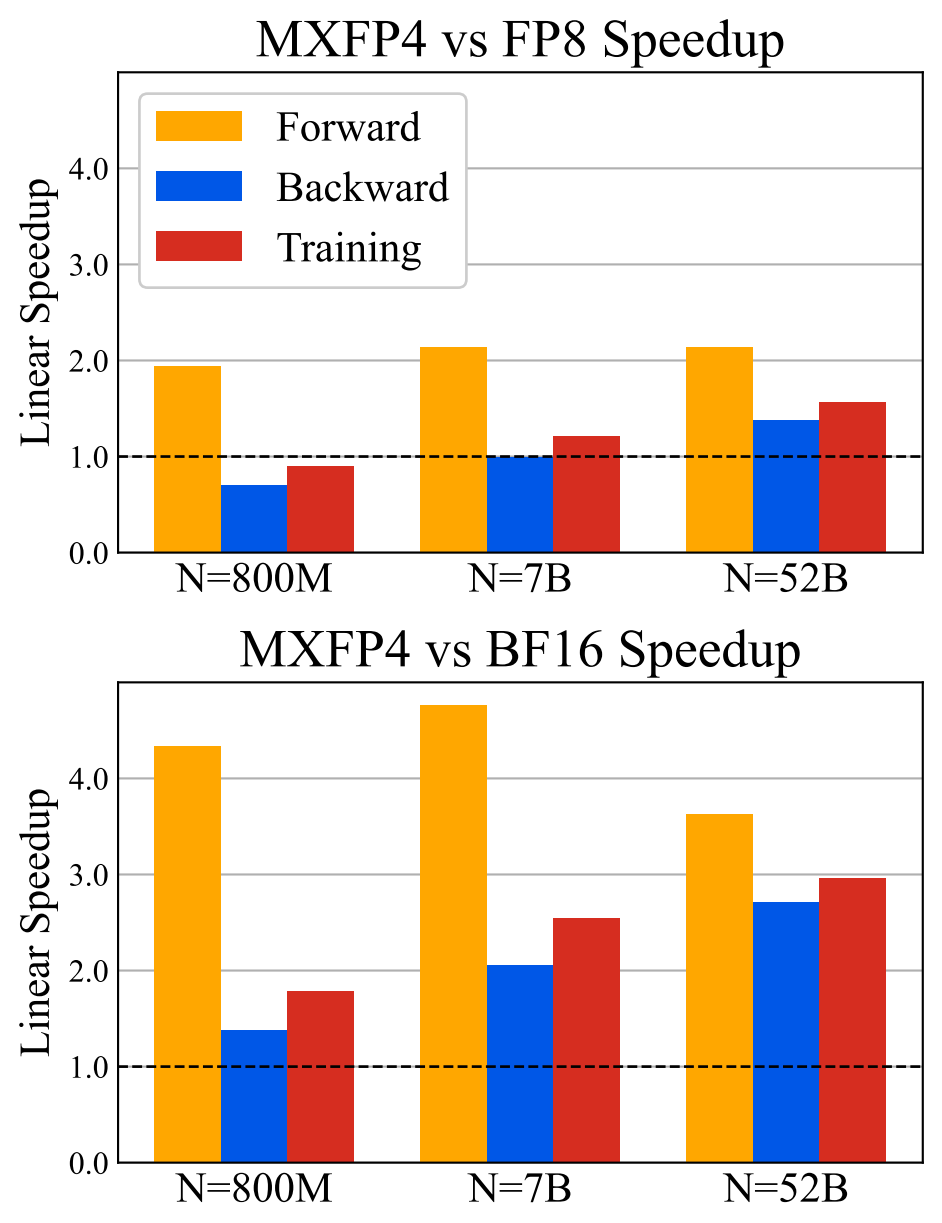
MXFP4 vs FP8 Speedup
Forward
Backward
Training
Linear Speedup
4.0
3.0
2.0
1.0
0.0
N=800M
N=7B
N=52B
MXFP4 vs BF16 Speedup
Linear Speedup
4.0
3.0
2.0
1.0
0.0
N=800M
N=7B
N=52B