Quality-Compression Pareto Frontier

- 3-bit Uniform
- INT4 (AWQ)
- 3.5-bit (Ours)
- INT8
- FP16

← Lower Error

Higher Compression →

X-axis: **Compression Ratio (vs FP16)**
Y-axis: **Normalized RMSE (%)**