AdamW Optimal Learning Rate Scaling Law
(ω = 4.0, Top-K = 5, Dataset = fineweb_100)

Fit 1 (Non-emb): $8.60e+01 \times P^{-0.635}$
Fit 2 (Compute): $2.51e+01 \times C^{-0.357}$

L=15,H=20
LR=3.80e-04

L=18,H=24
LR=2.68e-04

L=24,H=32
LR=1.55e-04

L=30,H=40
LR=1.01e-04

L=15,H=20
LR=3.80e-04

L=18,H=24
LR=2.67e-04

L=24,H=32
LR=1.50e-04

L=30,H=40
LR=9.51e-05

Point size $\propto$ weight
(best LR has largest weight)

Predictions use H=(4/3)L, Q=64

Learning Rate (LR)

Non-embedding Parameters