AW Optimal LR Scaling Law (Direct Weighted Fit)
(Target ω = 4.0, Top-K = 3, Dataset = fineweb_100)

Point size ∝ weight
(best LR has largest weight)

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

Fit 1 (Non-emb): $4.78e+01 \times P^{-0.600}$
Fit 3 (Compute): $1.36e+01 \times C^{-0.334}$

L=15,H=20
LR=4.21e-04
(Fit 3)
LR=4.19e-04
(Fit 1)

L=18,H=24
LR=3.02e-04
(Fit 3)
LR=3.02e-04
(Fit 1)

L=24,H=32
LR=1.77e-04
(Fit 3)
LR=1.80e-04
(Fit 1)

L=30,H=40
LR=1.15e-04
(Fit 3)
(Fit 1)

Learning Rate (LR)

Non-embedding Parameters