Adam-Star Optimal Learning Rate Scaling Law
(ω = 4.0, Top-K = 3, Dataset = fineweb_100)

Fit 1 (Non-emb): $1.29\text{e}{+}00 \times P^{-0.417}$
Fit 2 (Compute): $5.07\text{e}{-}01 \times C^{-0.230}$

L=15,H=20
LR=3.90e-04

L=18,H=24
LR=3.10e-04

L=24,H=32
LR=2.16e-04

L=30,H=40
LR=1.64e-04

L=15,H=20
LR=3.97e-04

L=18,H=24
LR=3.16e-04

L=24,H=32
LR=2.18e-04

L=30,H=40
LR=1.63e-04

Point size ∝ weight
(best LR has largest weight)

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

Learning Rate (LR)

Non-embedding Parameters