Best Validation Loss vs Compute (Non-embedding)
ademamix | eryngii_sweeps

Legend:
- Depth 2, 4 heads
- Depth 3, 5 heads
- Depth 4, 6 heads
- Depth 6, 7 heads
- Depth 8, 8 heads
- Depth 10, 9 heads
- Depth 12, 10 heads
- No offset: $2.07e+01 \times C^{-0.044}$ (R^2=1.000)
- With offset: $8.60e-06 + 2.08e+01 \times C^{-0.044}$ (R^2=1.000)

X-axis: Compute (FLOPs)
Y-axis: Best Validation Loss