Best Learning Rate vs Compute (Total)
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: $1.22\text{e}{+}05 \times C^{-0.450}$ (R^2=0.873)

X-axis: Compute (FLOPs)
Y-axis: Best Learning Rate