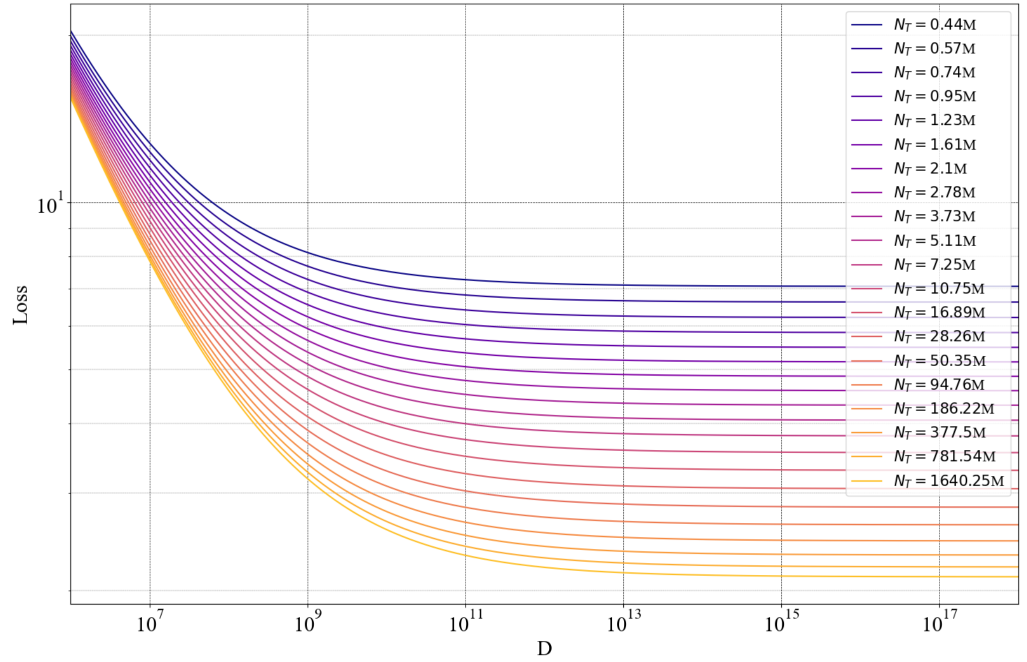
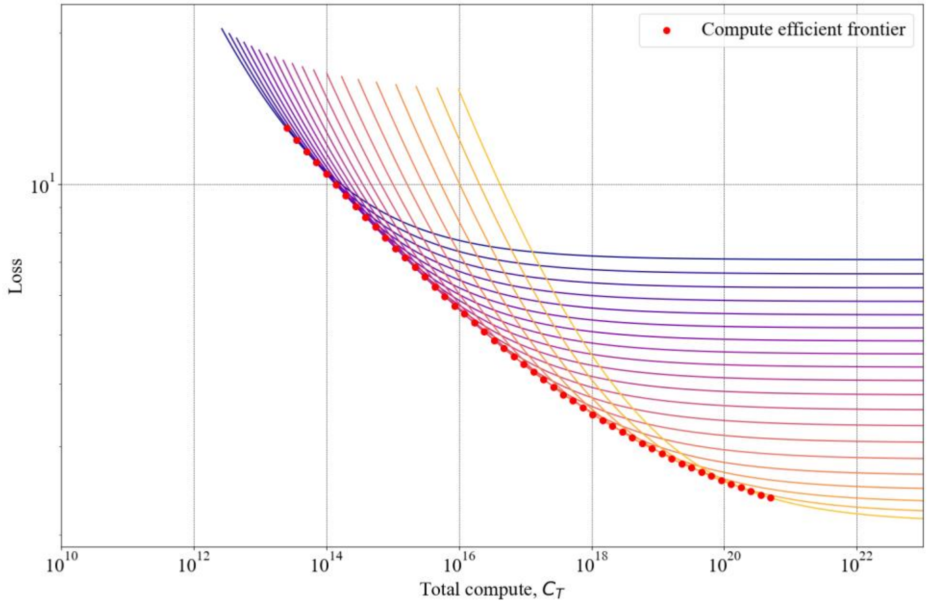
Start from fitted model of **Chinchilla's** training curves

$$\text{Loss}(N_T, D) = \frac{482}{N_T^{0.35}} + \frac{2085}{D^{0.37}} + 1.82$$

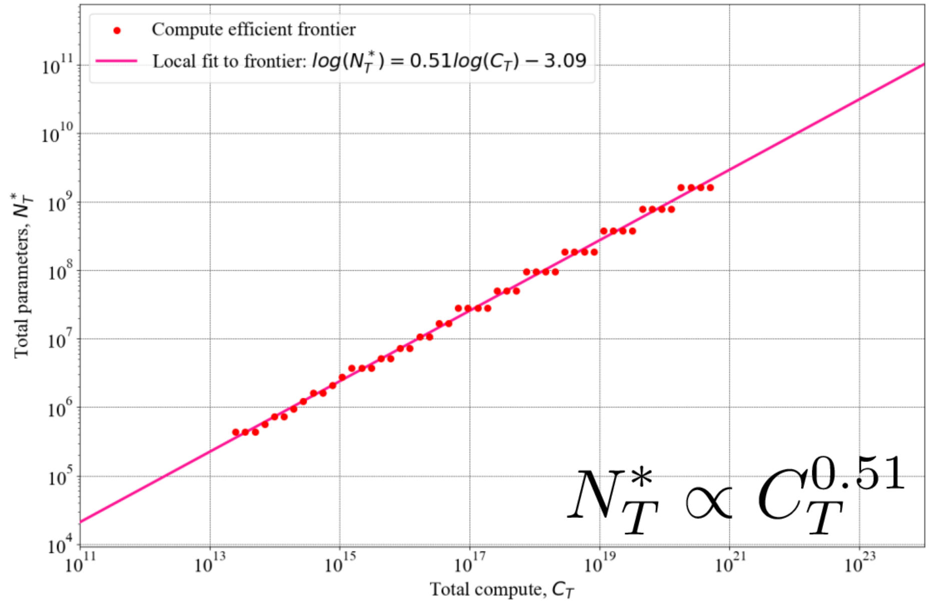
Generate training curves for model sizes used in **Kaplan's** study (1k to 1.5B params)



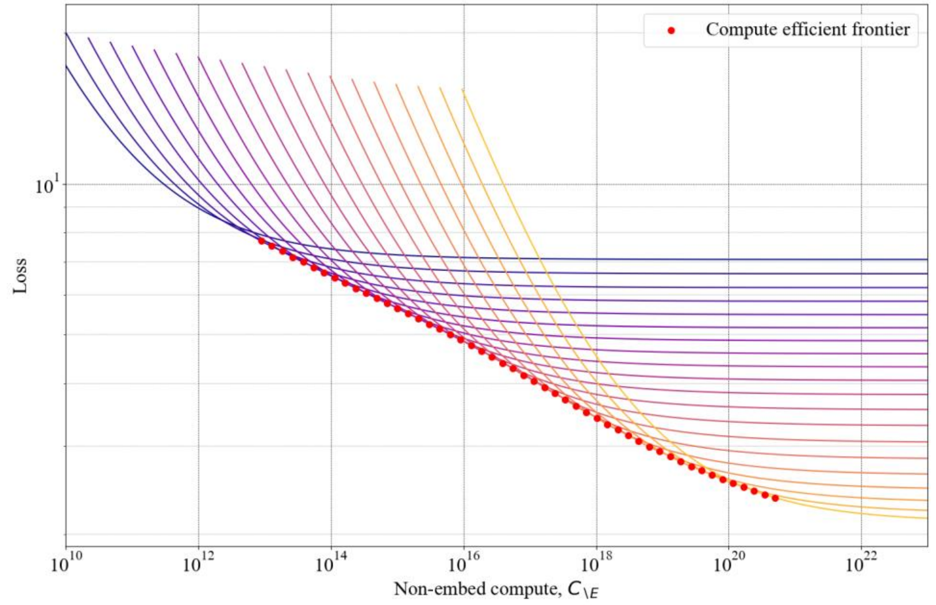
Find compute optimal frontier in terms of **total** parameters N_T and compute as in **Chinchilla**



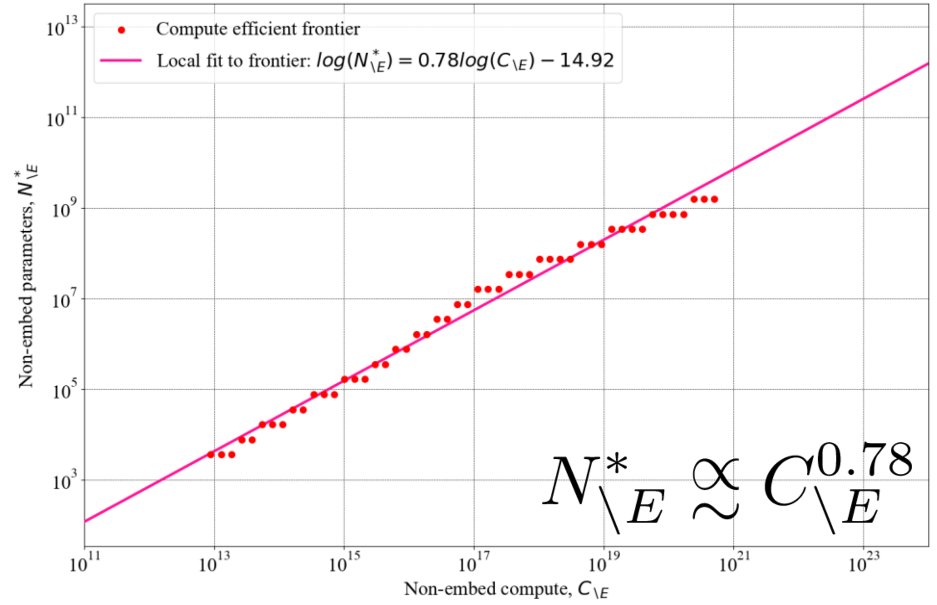
Find power law scaling coefficient of 0.51, close to Chinchilla's 0.50



Find compute optimal frontier in **non-embedding** parameters $N_{\setminus E}$ and compute as in **Kaplan**



Find **local** power law scaling coefficient of 0.78, close to Kaplan's 0.73


Compute efficient frontier
Local fit to frontier: $log(N_{\setminus E}^*) = 0.78 log(C_{\setminus E}) - 14.92$
Non-embed parameters, $N_{\setminus E}^*$
10^3
10^5
10^7
10^9
10^{11}
10^{13}
Non-embed compute, $C_{\setminus E}$
10^{11}
10^{13}
10^{15}
10^{17}
10^{19}
10^{21}
10^{23}
$N_{\setminus E}^* \underset{\sim}{\propto} C_{\setminus E}^{0.78}$