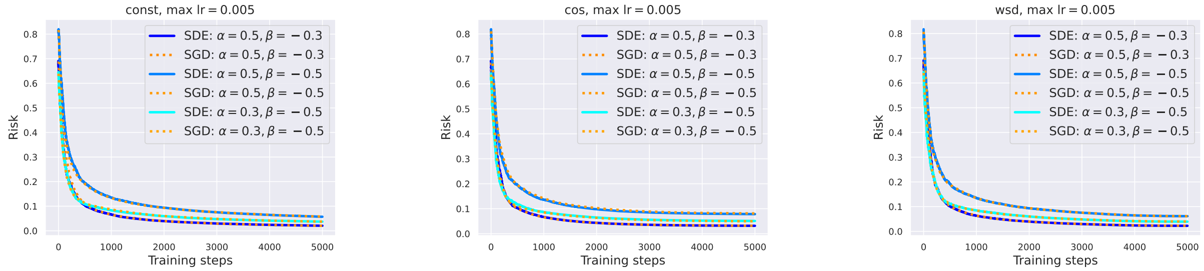


(a) Constant LRS

(b) Cosine LRS

(c) WSD LRS

Figure 1: SDE and SGD loss comparison under our linear regression setup with different LRS and α, β's

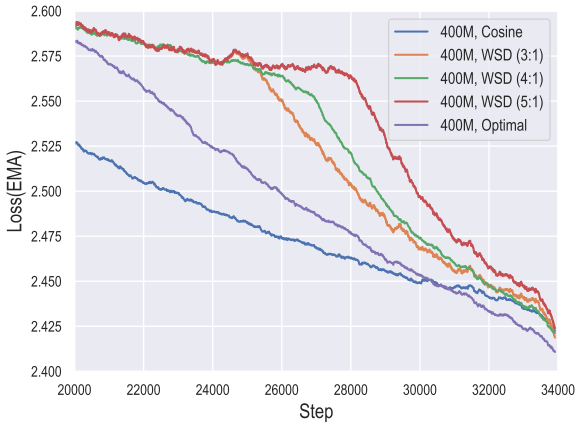
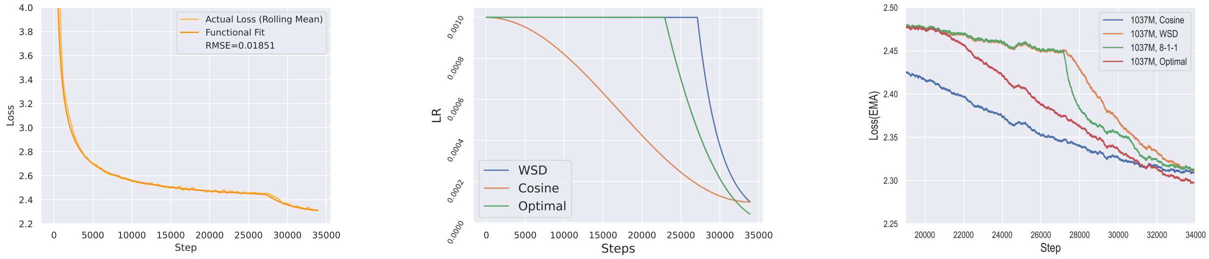


Figure 2: 400M Llama (dense) model, 20B data. WSD with different decaying ratio compared with the optimal

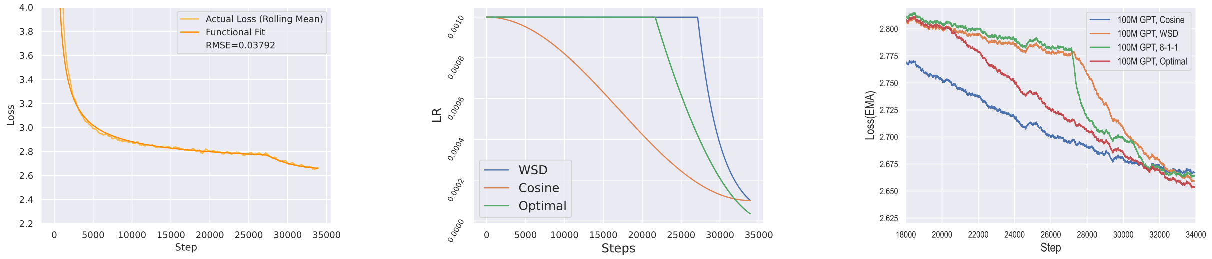


(a) Fitting result

(b) Optimal LRS

(c) Loss curves

Figure 3: 1B Llama (dense) model, 20B data

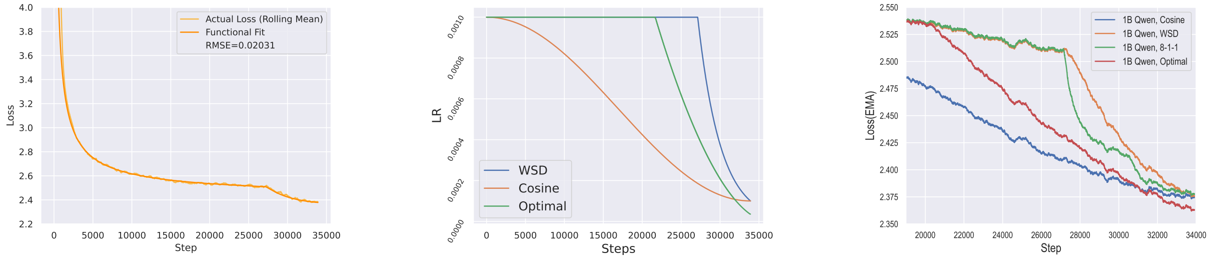


(a) Fitting result

(b) Optimal learning rate

(c) Loss curves

Figure 4: 100M GPT2 (dense) model, 20B data

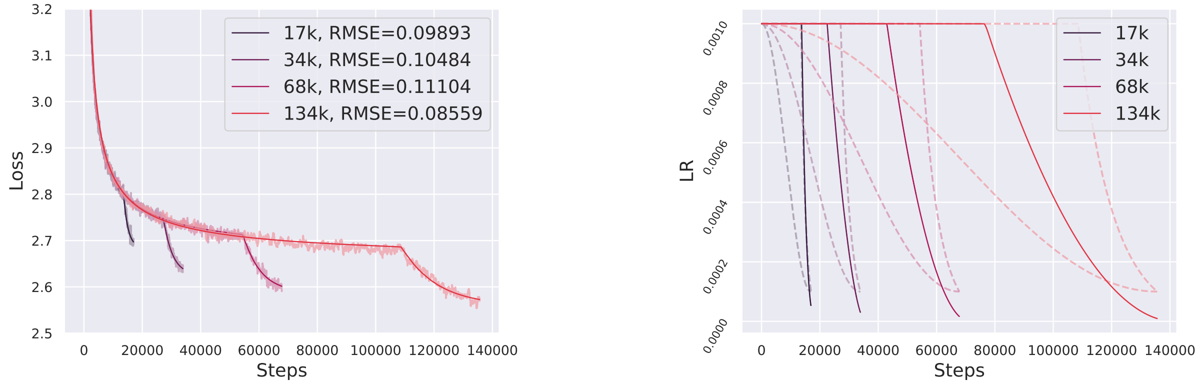


(a) Fitting result

(b) Optimal learning rate

(c) Loss curves

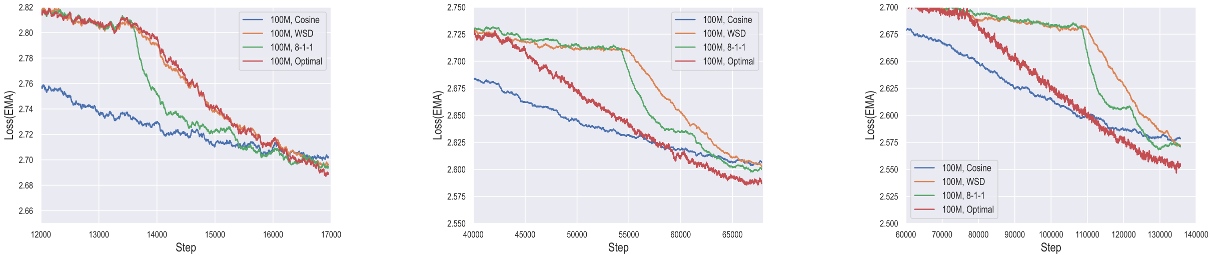
Figure 5: 1B Qwen (MoE) model, 20B data



(a) Fitting result

(b) Optimal learning rate

Figure 6: Fitted functional scaling laws and optimal LRS on different total training steps



(a) 17k steps

(b) 68k steps

(c) 134k steps

Figure 7: Loss curves on different total training steps

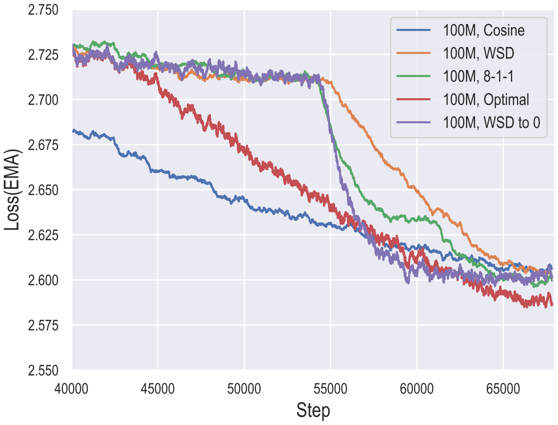


Figure 8: 100M Llama (dense) model, 40B data. WSD decaying to zero does not bring significant loss reduction