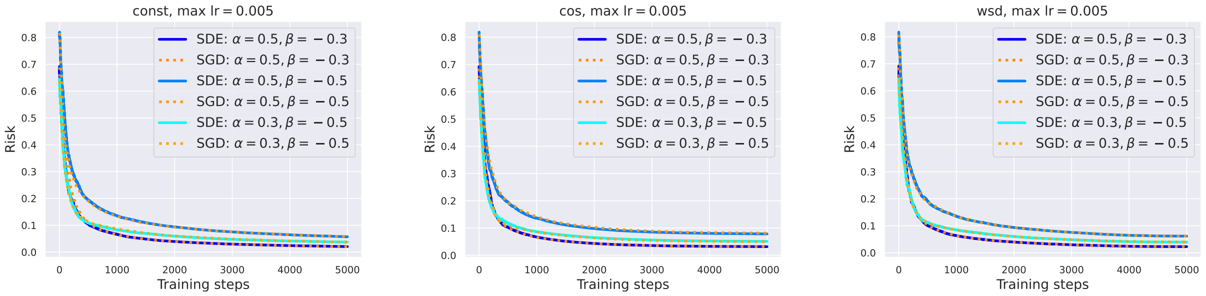


(a) Constant LRS

(b) Cosine LRS

(c) WSD LRS

Figure 1: **Empirical Validation of Our SDE Modeling of SGD:** In our linear regression setup ($M = 1000, S = 5000, B = 4, \sigma = 0.1$), we train the student model with model size $N = 64$ using two algorithms, SGD and SDE, across various LRS settings with the same maximum learning rate $\eta_{\max} = 5 \times 10^{-3}$ (including constant LRS, cosine LRS, and WSD LRS) under different α and β regimes. The SGD curve represents the population risk at each training step (i.e., $\mathcal{R}(\boldsymbol{v}_r)$), while the SDE curve represents the population risk of the SDE at the corresponding intrinsic time (i.e., $\mathcal{R}(\tilde{\boldsymbol{v}}_{T(r)})$). All experiments show that, under our setting, the SDE model and the corresponding SGD exhibit very similar behavior.

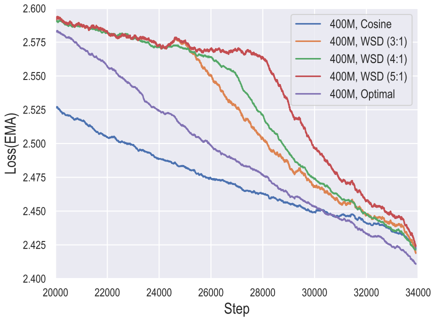
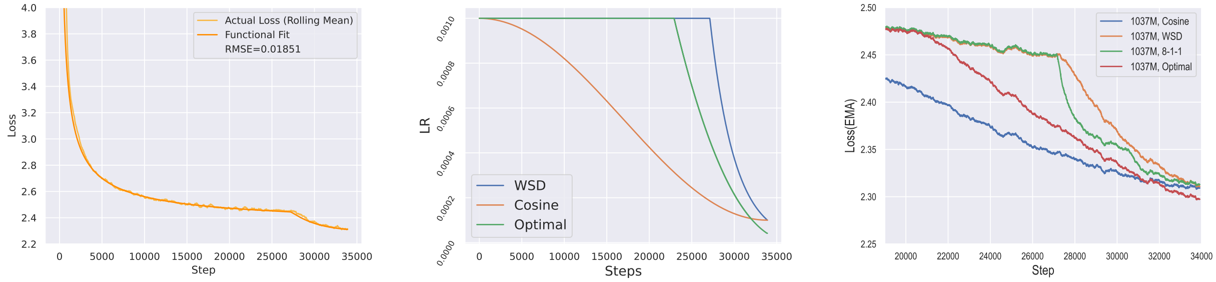


Figure 2: **WSD with Different Decay Times:** We train a 400M LLaMA (dense) model with 20B tokens of training data and WSD LRSs with the ratios between stable time and decay time of 3:1, 4:1, and 5:1. All WSD LRSs exhibit a final loss similar to that of the Cosine LRS, and the optimal LRS derived from our functional scaling law outperforms all other LRSs by a loss gap of approximately 0.01.

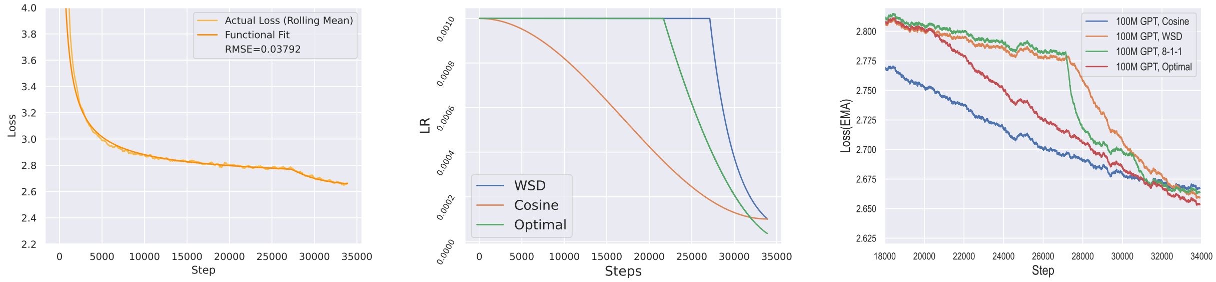


(a) Fitting on WSD LRS

(b) Proposed and Existing LRSs

(c) Loss curve of LRSs

Figure 3: **Experiment on the 1B LLaMA (Dense) Model.** Figure (a): We fit our functional scaling law on the loss curve of 1B LLaMA (dense) model with 20B tokens training data and WSD LRS. Figures (b)(c): The comparison on the 1B model between the "optimal" LRS, cosine LRS, WSD LRS with exponential decay and "8-1-1" LRS.

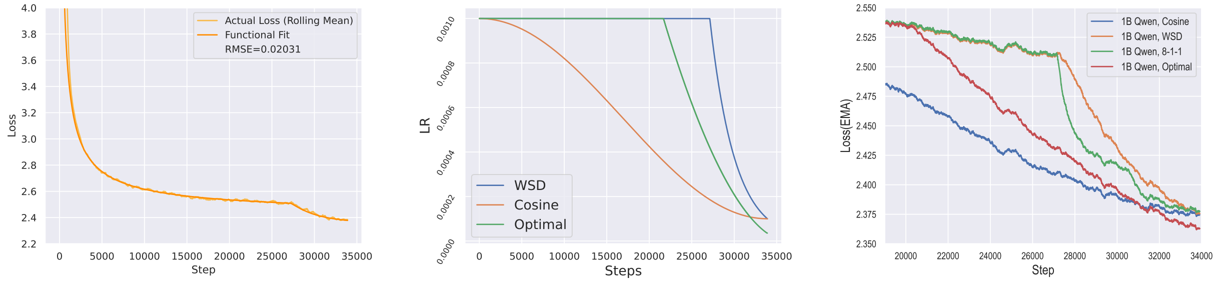


(a) Fitting on WSD LRS

(b) Proposed and Existing LRSs

(c) Loss curve of LRSs

Figure 4: **Experiment on the 100M GPT2 (Dense) Model.** Figure (a): We fit our functional scaling law on the loss curve of 100M GPT2 (dense) model with 20B tokens training data and WSD LRS. Figures (b)(c): The comparison on the 100M model between the "optimal" LRS, cosine LRS, WSD LRS with exponential decay and "8-1-1" LRS.

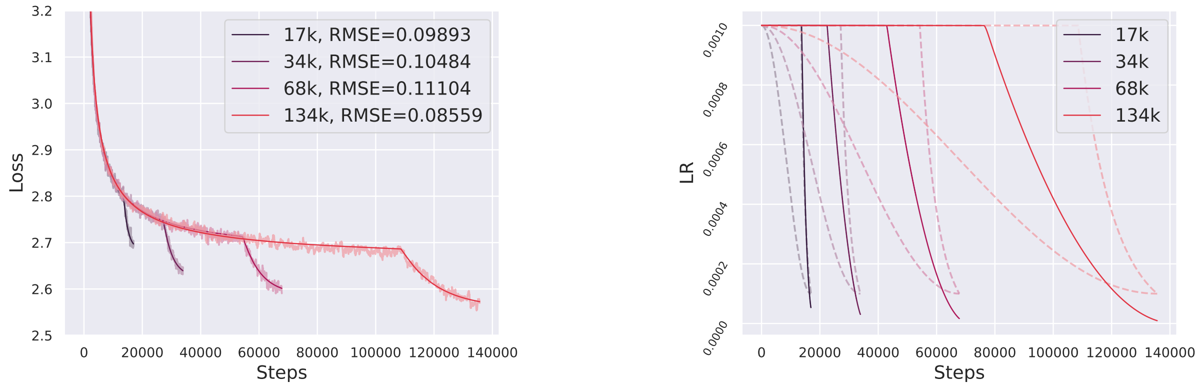


(a) Fitting on WSD LRS

(b) Proposed and Existing LRSs

(c) Loss curve of LRSs

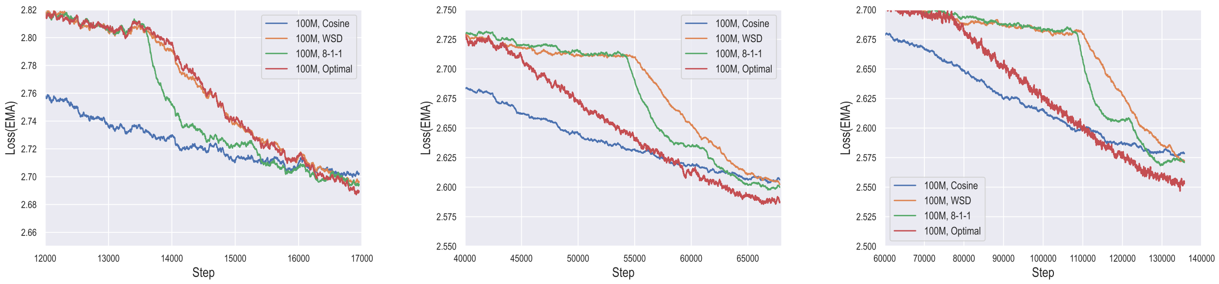
Figure 5: **Experiment on the 1B Qwen (MoE) Model.** Figure (a): We fit our functional scaling law on the loss curve of 1B Qwen (MoE with 230M activated parameters) model with 20B tokens training data and WSD LRS. Figures (b)(c): The comparison on the 1B model between the "optimal" LRS, cosine LRS, WSD LRS with exponential decay and "8-1-1" LRS.



(a) Fitting result

(b) Optimal LRS

Figure 6: **Experiments with Different Total Steps.** Figure (a): Fitted functional scaling laws on 100M LLaMA model with different total training steps 17k, 34k (original result), 68k and 134k. Figure (b): Optimal LRSs compared with cosine and WSD LRSs. The solid lines are optimal LRSs, and the dashed lines are cosine/WSD LRSs.



(a) 17k steps

(b) 68k steps

(c) 134k steps

Figure 7: **Experiments with Different Total Steps.** We compare loss curves of existing LRSs and "optimal" LRS on the 100M LLaMA model with different total training steps 17k, 68k and 134k.

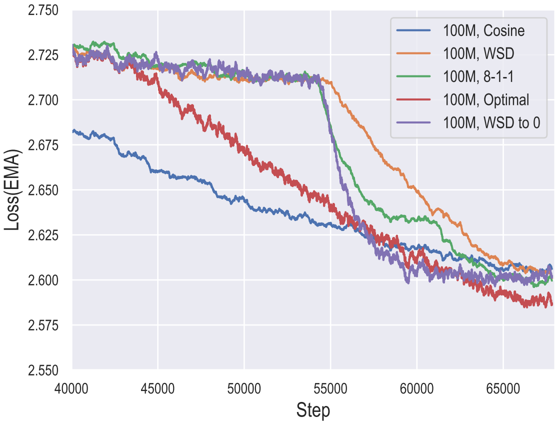


Figure 8: **Comparison Between "Optimal" LRS and WSD with a Near-Zero Final LR:** We train a 100M LLaMA (dense) model with 40B tokens training data and various LRSs with the same $LR_{\max} = 10^{-3}$, including WSD LRS with $LR_{\min} = \frac{1}{10}LR_{\max}$, WSD LRS with $LR_{\min} = 10^{-7}$, cosine LRS with $LR_{\min} = \frac{1}{10}LR_{\max}$, "8-1-1" LRS with $LR_{\min} = \frac{1}{10}LR_{\max}$, and "optimal" LRS. The experimental results show that decaying to (near) zero does not result in significant loss reduction.