Moonlight Intermediate Checkpoints

1. How to download

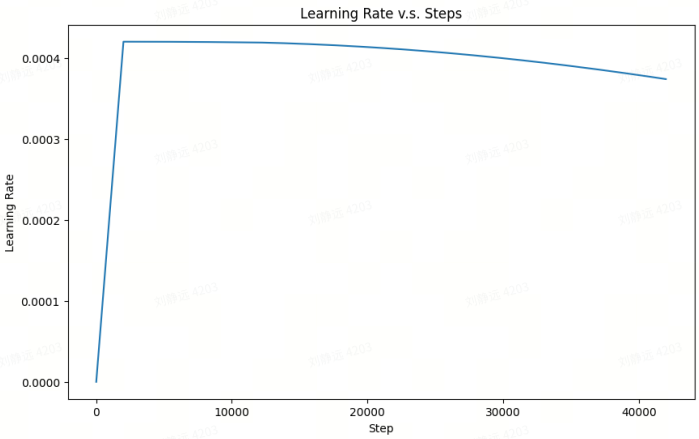
https://drive.google.com/drive/folders/1j-zOOEP8sUpCVzI5IphLM3jkhFXXIokl

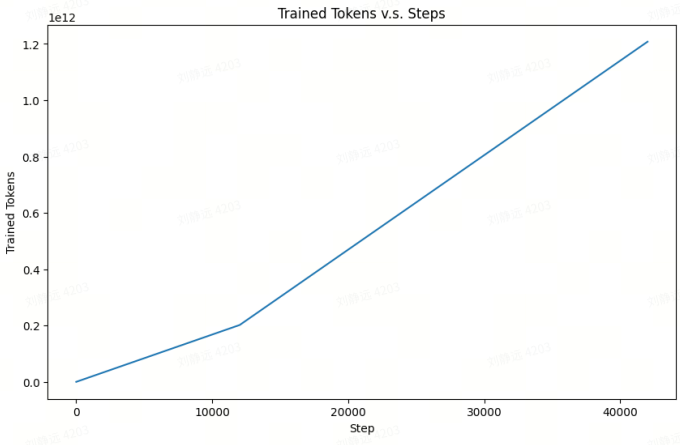
- Moonlight_hf_step_XX: the Moonlight checkpoints trained by Muon at step XX
- Moonlight_adam_hf_step_XX: the Moonlight-A checkpoints trained by AdamW at step XX

2. Training Schedule

Both Moonlight and Moonlight-A are trained with the same data in the same training schedule:

1. Warmup: the LR linear warm-ups in 2K steps to its peak 4.2e-4, with global batch size of 2048 examples;
2. Cosine decay: the LR starts a cosine decay schedule from 4.2e-4 to 4.2e-5 in 5.2T tokens
    a. From 2K to 12K steps, the model used a global batch size of 2048 examples;
    b. From 12K to 42K steps, the model used a global batch size of 4096 examples;

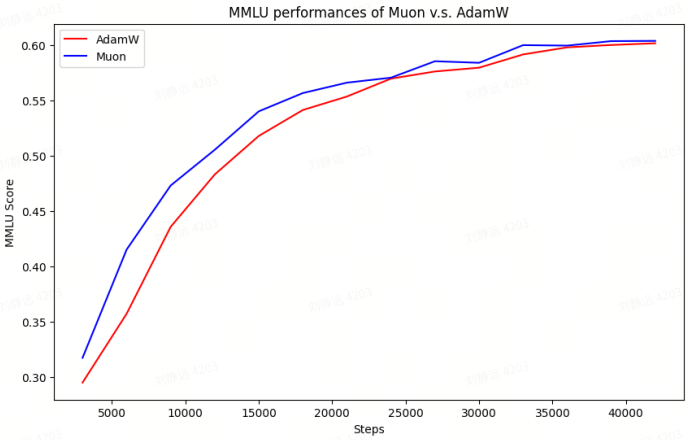


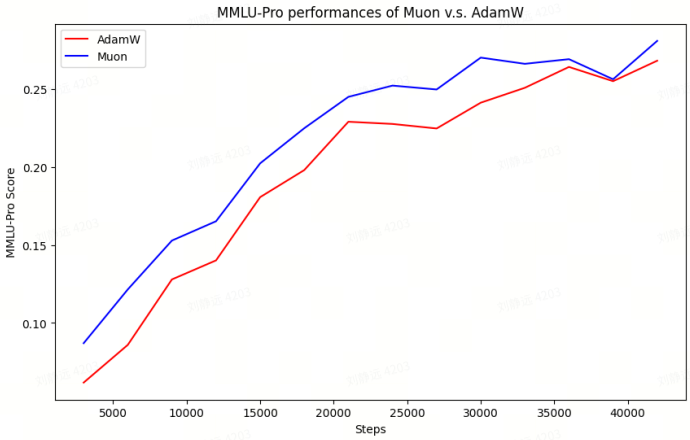


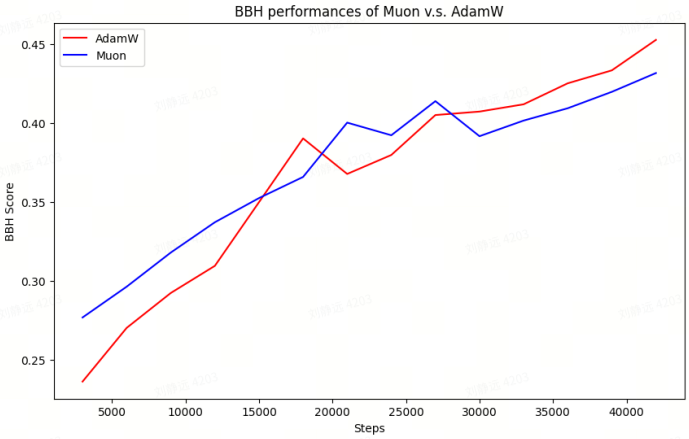
3. Performances

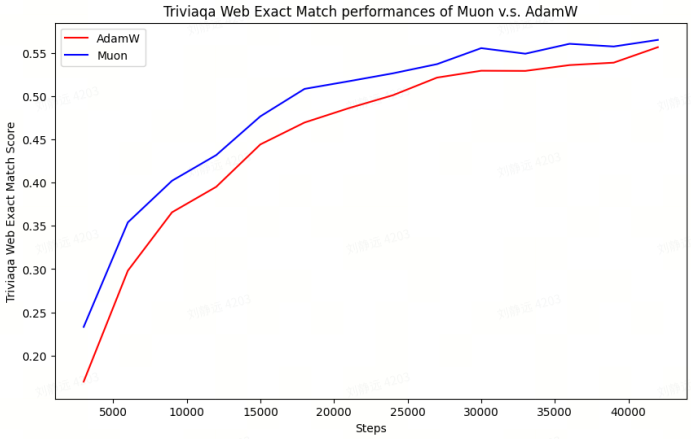
We compared the model's performances in terms of 4 categories as described in our paper. As discussed in the paper, we observed that Muon especially performed better on **Code and Math** related bmks.

- **English Language Understanding and Reasoning:** MMLU, MMLU-pro, BBH, TriviaQA

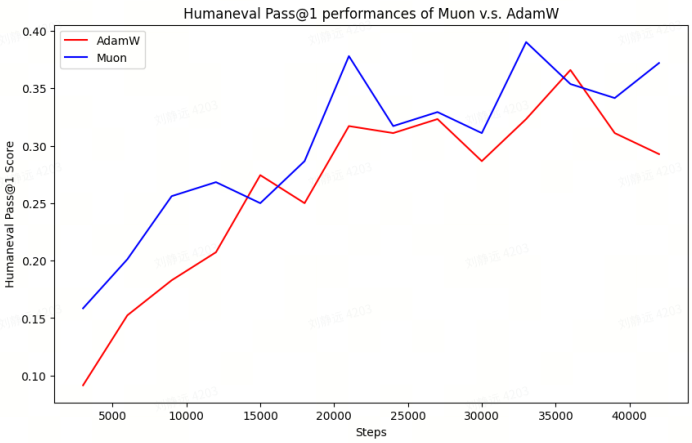


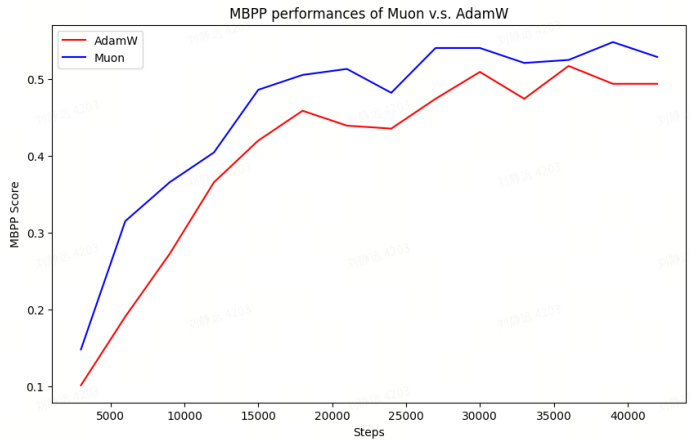




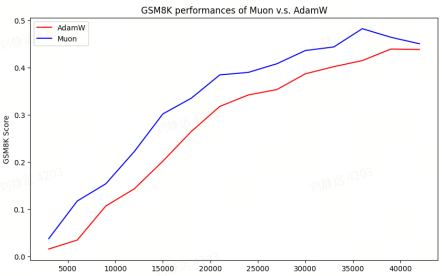


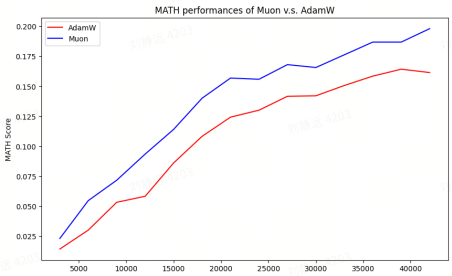
- **Code Generation:** HumanEval, MBPP

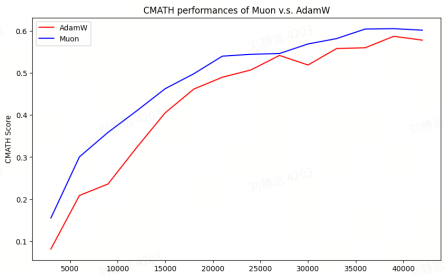




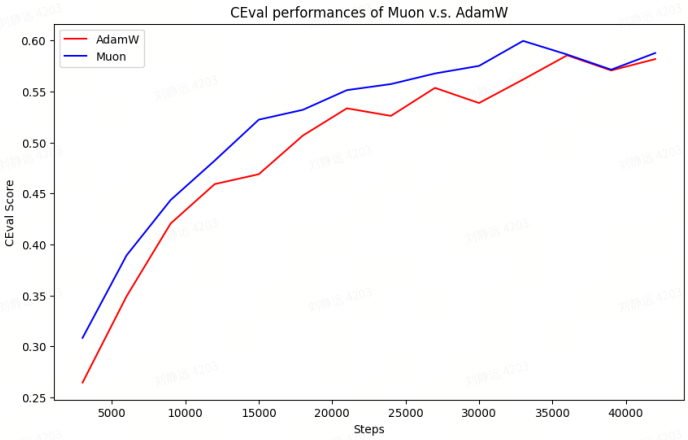
- **Mathematical Reasoning:** GSM8K, MATH, CMATH

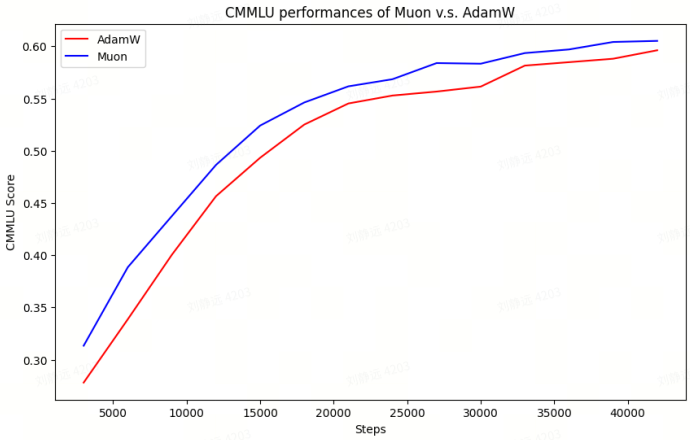






- **Chinese Language Understanding and Reasoning:** C-Eval, C-MMLU





4. Dynamics of Singular Spectrum and Srank

We showed the conclusion of Sec. 3.5 as discussed in our paper.

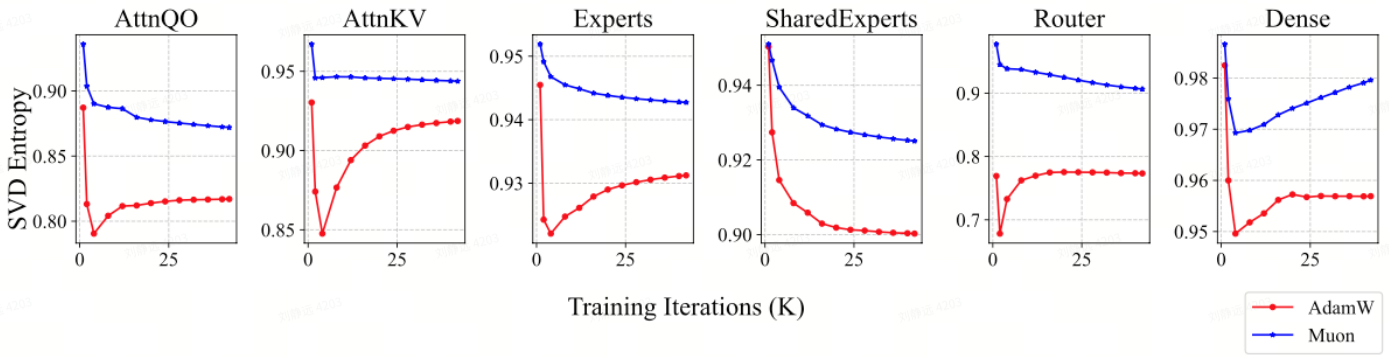


Figure 4: SVD entropy of weight matrices across different training iterations. We categorize the weight matrices into 6 different groups: 1) AttnQO denotes the weight matrices related to the query and output projection in the attention layer; 2) AttnKV denotes the weight matrices related to the key and value projection in the attention layer; 3) Experts denotes the weight matrices in expert models; 4) SharedExperts denotes the weight matrices in shared expert models; 5) Router denotes the weight matrices in the router; 6) Dense denotes the weight matrices in the first dense layer. The SVD entropy is calculated as the macro-average of the weight matrices in each group across all layers. For weights in expert models, we only calculate 3 out of 64 experts in different layers for efficiency.

Besides, we also ran the 'srank' analyais on these intermediate ckpts and observed huge differences: