Paper 22

Franz Srambical

August 15, 2024

We investigate the impact of learning rate warmup on GPT-style Transformers using muP/SP trained on a *realistic* repository on language modeling. We train on `wikitext-2` for a single epoch and report the validation loss.

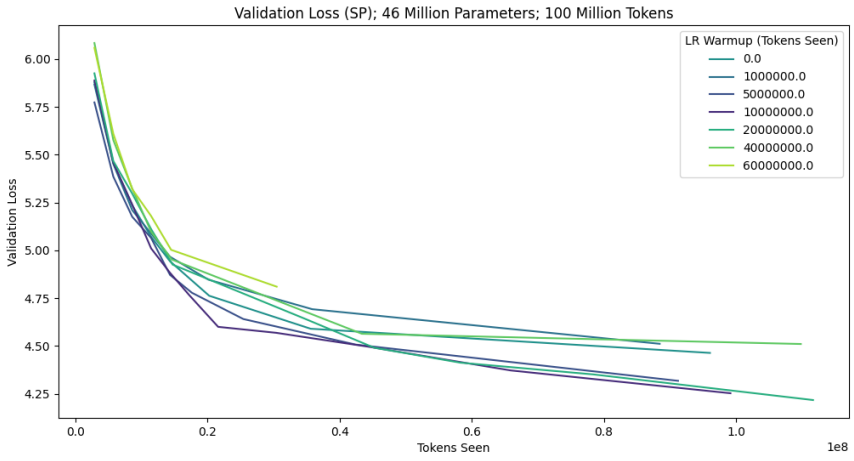
In the extension phase, we wanted to answer the following questions: 1) Does muP lead to performance gains in *practical* settings? 2) Do we still need learning rate warmup under muP?

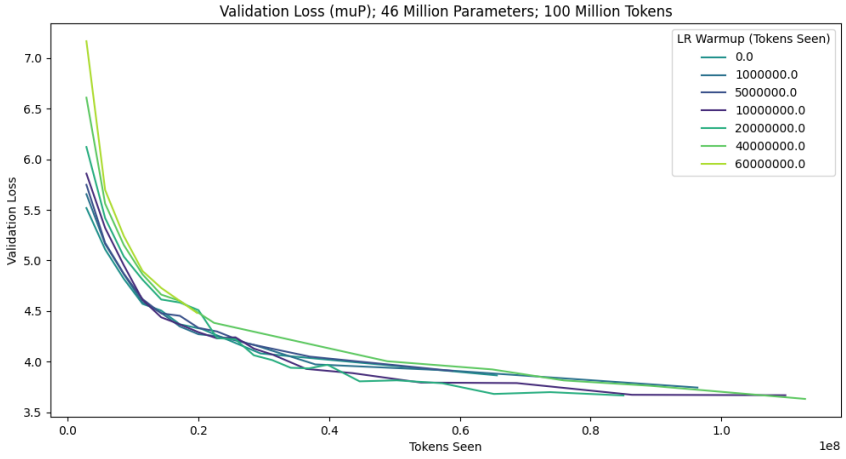
We are in the *realistic* setting since we are using hlb-gpt (Hyperlightspeedbench-gpt [1]), which incorporates numerous techniques that are used in the large-scale setting by all major research labs in industry. These include fused `attention+MLP` blocks, a dynamic microbatch scheduler based on the expected gradient norm, parameter-group dependent learning rates and schedules, an empirically motivated scaling of the learning rate with model size as well as sequence length warmup with maximum batch size calculations based on the available VRAM.

Since latter technique leads to 'unaligned' loss curves, we modify the repository such that learning rate schedule and warmup are based on `tokens_seen` instead of the actual step count. That way we ensure that loss curves across runs are aligned. For the muP experiments, we remove the parameter-group dependent learning rates and schedules as well as the empirically motivated learning rate scaling factors from the repository, since these are techniques that muP natively implements.

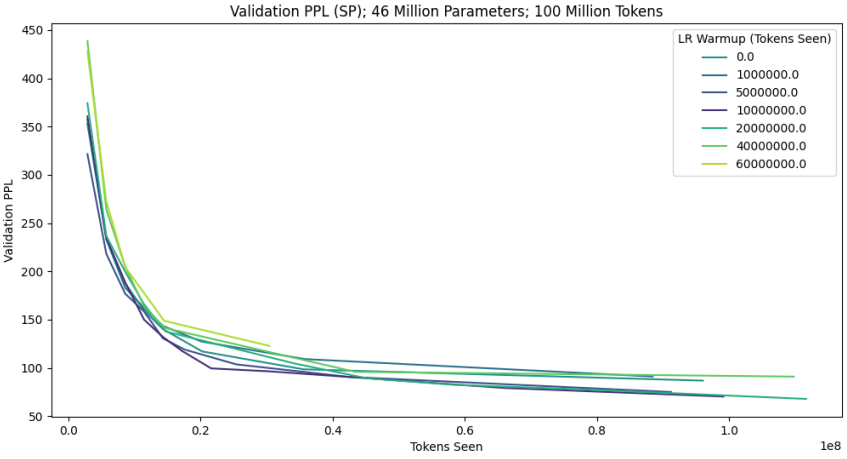
[1] https://github.com/tysam-code/hlb-gpt/

Our experiments yield the following validation loss and perplexity curves:

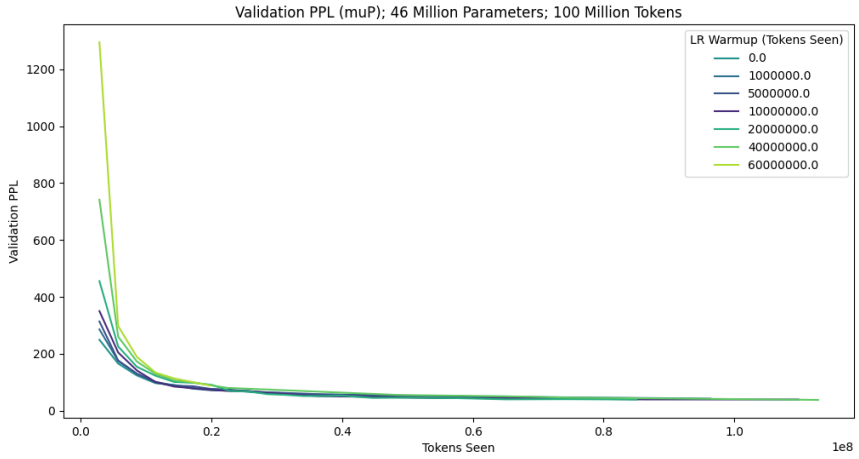



Validation Loss (muP); 46 Million Parameters; 100 Million Tokens
Validation Loss
Tokens Seen
1e8
LR Warmup (Tokens Seen)
0.0
1000000.0
5000000.0
10000000.0
20000000.0
40000000.0
60000000.0

24


Validation PPL (SP); 46 Million Parameters; 100 Million Tokens
Validation PPL
Tokens Seen
1e8
LR Warmup (Tokens Seen)
0.0
1000000.0
5000000.0
10000000.0
20000000.0
40000000.0
60000000.0

25

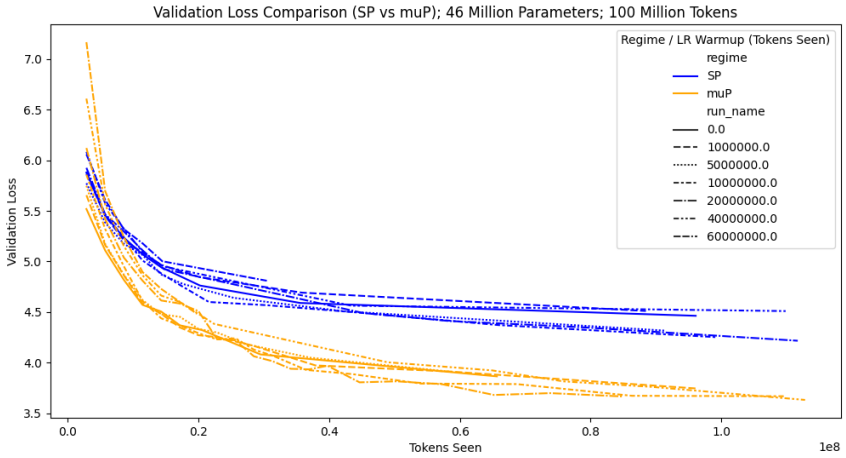

Validation PPL (muP); 46 Million Parameters; 100 Million Tokens
Validation PPL
Tokens Seen
1e8
LR Warmup (Tokens Seen)
0.0
1000000.0
5000000.0
10000000.0
20000000.0
40000000.0
60000000.0

26

Here, the color difference between a curve and the best curve depicts the distance from the best run in terms of learning rate warmup. The plots depict a clear color gradient, i.e. the farther a run departs from the optimal learning warmup, the worse the convergence. Thus, learning rate warmup is still *helpful*, even under muP. However, the curves under muP are closer together than under SP, suggesting that learning rate warmup is more impactful under SP than under muP, answering our second question for the extension phase and confirming a hypothesis of the research community [2].

[2] https://cloneofsimo.notion.site/What-to-do-to-scale-up-09e469d7c3444d6a90305397c38a46f5

Lastly, we plot the validation loss curves of muP and SP in the same plot:



Here, we see that muP consistently outperforms SP in our experiments, affirmatively answering our first question of the extension phase. Not only that, but muP outperforms SP in a repository that implements empirical scaling laws based learning rates as well as parameter group dependent initialization, learning rates and schedules.