1 Additional experiments

We additionally conducted two relevant experiments: the application of weight decay scheduling strategies with the AdamW optimizer, and Dependent t-test for paired samples.

1.1 LLM Pre-training with AdamW

Table 1 provides a comparison of several weight decay scheduling strategies for pre-training LLaMa-60M and LLaMa-130M models with the AdamW optimizer. The results clearly demonstrate the effectiveness of applying weight decay, as all scheduling strategies outperform the baseline with no weight decay (WD=0) in terms of validation perplexity.

Table 1: **(AdamW.)** Comparison of various weight decay scheduling strategies using AdamW optimizer for pre-training LLaMa-60M and LLaMa-130M models under different weight decay values. Validation perplexity (↓) on the C4 dataset is reported. All baselines are carefully tuned. 'WD=0' indicates that weight decay is disabled during model training.

	LLaMa-60M			LLaMa-135M		
Weight Decay	0.1	0.05	0.01	0.1	0.05	0.01
WD=0		32.73			24.39	
`Uniform`	31.95	32.31	32.66	23.32	23.81	24.28
`AWD`	32.58	32.67	32.67	24.30	24.35	24.41
`Adadecay`	31.88	32.25	32.58	23.18	23.62	24.21
`AlphaDecay`	31.20	31.65	32.45	22.66	23.04	23.98

`AlphaDecay` consistently outperforms other weight decay scheduling strategies across different model sizes and hyperparameter settings, demonstrating superior regularization and generalization when training with AdamW. These results highlight the robustness and effectiveness of `AlphaDecay`, supporting its adoption for optimizing large-scale transformer-based language models.

1.2 Dependent t-test for paired samples

Table 2 provides a comparison of several weight decay scheduling strategies using the Adam optimizer, evaluated through repeated experiments with different random seeds.

Table 2: **(Dependent t-test with Adam.)** Each method (`Uniform`, `AWD`, `AdaDecay`, and `AlphaDecay`) was evaluated by conducting six repeated experiments with random seeds { 5, 6, 7, 8, 9, 10 }. Validation perplexity is reported as mean ± standard deviation. For each weight decay setting, a dependent t-test for paired samples was performed, comparing `AlphaDecay` against `Uniform`, `AWD`, and `AdaDecay`, respectively. The resulting p-values are presented alongside perplexity scores.

Method	Weight Decay=0	Weight Decay=1e-5		Weight Decay=5e-6		Weight Decay=1e-6	
	Perplexity	Perplexity	P-value	Perplexity	P-value	Perplexity	P-value
`Uniform`	24.55 ± 0.07	22.97 ± 0.07	8.38e-4	23.12 ± 0.03	1.47e-6	24.12 ± 0.04	1.05e-7
`AWD`		24.13 ± 0.15	6.94e-6	24.46 ± 0.07	5.34e-8	24.53 ± 0.03	5.34e-9
`AdaDecay`		23.18 ± 0.05	1.46e-5	23.07 ± 0.04	1.11e-7	24.00 ± 0.03	2.69e-9
`AlphaDecay`		22.77 ± 0.02		22.54 ± 0.03		23.44 ± 0.02	

The results demonstrate the benefit of applying weight decay for improved validation perplexity, with `AlphaDecay` consistently exhibiting superior performance and stability across all tested settings. The dependent t-test results further substantiate these findings, with statistically significant p-values supporting the advantage of `AlphaDecay` over `Uniform`, `AWD`, and `AdaDecay` in nearly all cases.