Pretraining	Cont training	arxiv	book	c4	cc	github	se	wiki	Avg PPL ↓
1.4B models									
Transformer	Transformer	4.54	12.87	19.21	13.77	3.99	9.49	10.96	10.69
Transformer	MTA	4.51	12.78	19.09	13.67	3.95	9.41	10.87	10.61
MTA	MTA	**4.47**	**12.57**	**18.82**	**13.47**	**3.89**	**9.25**	**10.65**	**10.44**
Llama 3 herd									
Llama 3.2 1B	Llama 3.2 1B	4.54	14.60	18.14	13.40	3.92	8.84	10.25	10.53
Llama 3.2 1B	MTA	4.52	14.49	18.04	13.32	3.89	8.79	10.19	10.46
Llama 3.2 3B	Llama 3.2 3B	4.12	12.08	15.58	11.50	3.37	7.52	8.35	8.93
Llama 3.2 3B	MTA	4.11	12.03	25.51	11.46	3.35	7.48	8.31	8.89
Llama 3.1 8B	Llama 3.1 B	4.00	11.04	15.04	10.99	3.29	7.33	8.00	8.53
Llama 3.1 8B	MTA								

Table 10: Validation perplexity on SlimPajama dataset after continuous training for 10.5B tokens on our 1.4B models, and continuous training for 5.3B for Llama 3 models.

I Finetuning with MTA.

One natural question readers might ask is if MTA could be integrated into models that were already trained with standard attention? Would this require complete retraining, or could it be added through some form of adaptation? Architectually, MTA can be added as additional layers to already trained models, and weights can be updated with continual training. Since the main component of MTA is a convolution, we can initialize it to identity and insert to existing a Transformer layer. Such a modification will not change the output of the layer, so when we start training the model, it should maintain its performance. However, as the added convolution starts deviating from the identity state, it will allow the transformer to condition its attention on multiple keys and queries.

To understand how well pre-trained model can learn we perform preliminary experiments by finetuning our 1.4B model, as well as opensourced Llama models (Grattafiori et al., 2024). In all these experiments we used same finetuning setup similar to Section 4.3, but kept the context length at 2048 tokens.

Validation perplexity results are reported in Table 10. We observe that all models finetuned with MTA were not only able to incorporate new kernels, but outperform baselines in terms of perplexity.