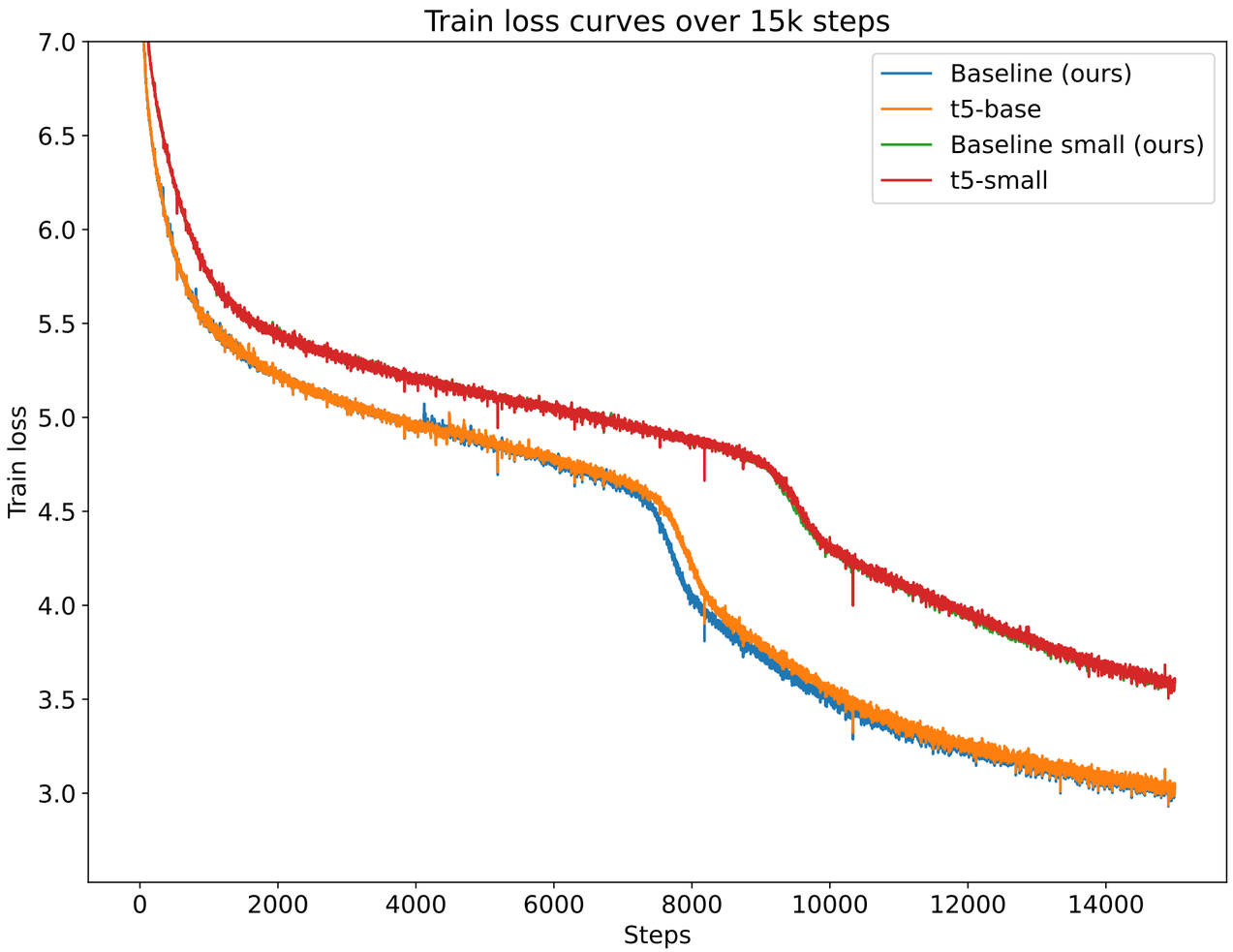
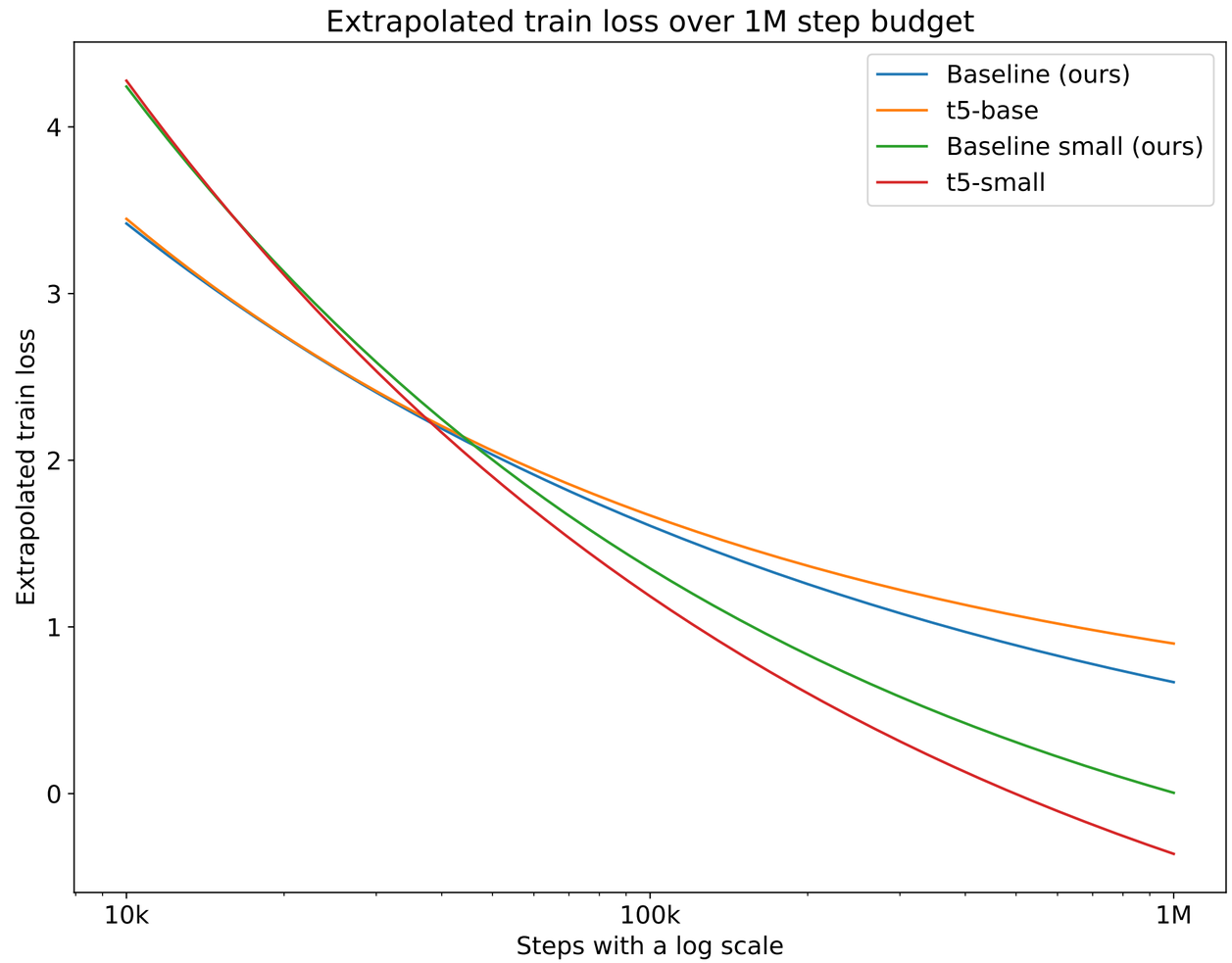
Our reimplementation and t5 baseline





Scaling law fit details and perplexity (PPL) predictions

Experiment	# params	Train - eval loss	Scaling law	PPL at 100k	PPL at 300k	PPL at 1M
Baseline (ours)	223M	0.140	52.42(t ** -0.286) - 0.34	4.987	2.952	1.951
t5-base	223M	0.141	90.03(t ** -0.365) - -0.32	5.305	3.395	2.460
Baseline small (ours)	61M	0.038	115.13(t ** -0.332) - 1.17	3.861	1.788	1.005
t5-small	61M	0.152	99.13(t ** -0.301) - 1.91	3.268	1.368	0.697