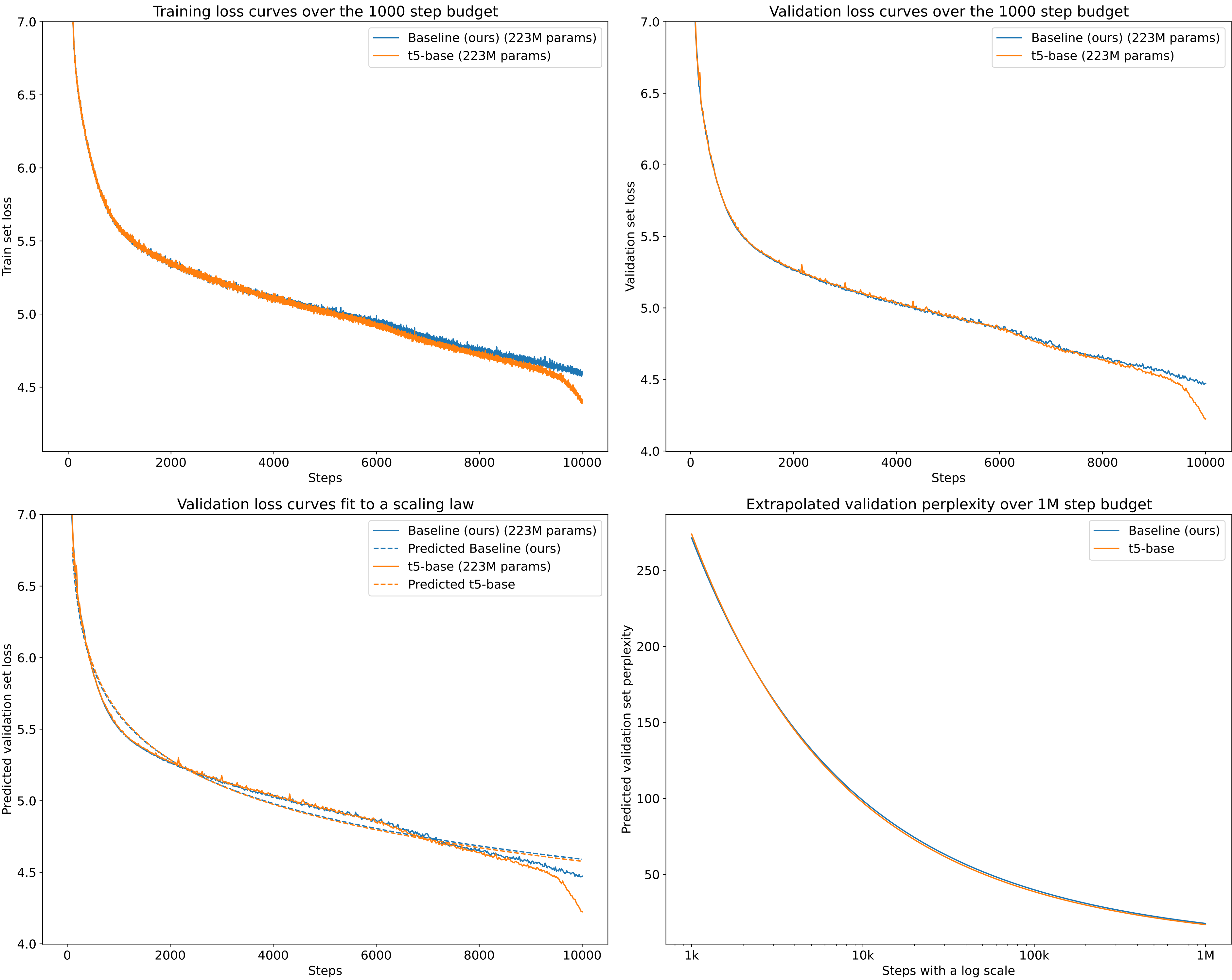


Our reimplementaion and t5 baseline



Scaling law fit details and perplexity (PPL) predictions

Experiment	Scaling law	Mean L1 residual	PPL at 1k	PPL at 10k	PPL at 100k	PPL at 300k	PPL at 1M
Baseline (ours)	$10.69(t^{**} - 0.049) - 3.96$	0.049	271.279	98.626	39.908	26.824	17.771
t5-base	$10.61(t^{**} - 0.050) - 3.84$	0.061	273.843	97.226	38.668	25.826	17.014