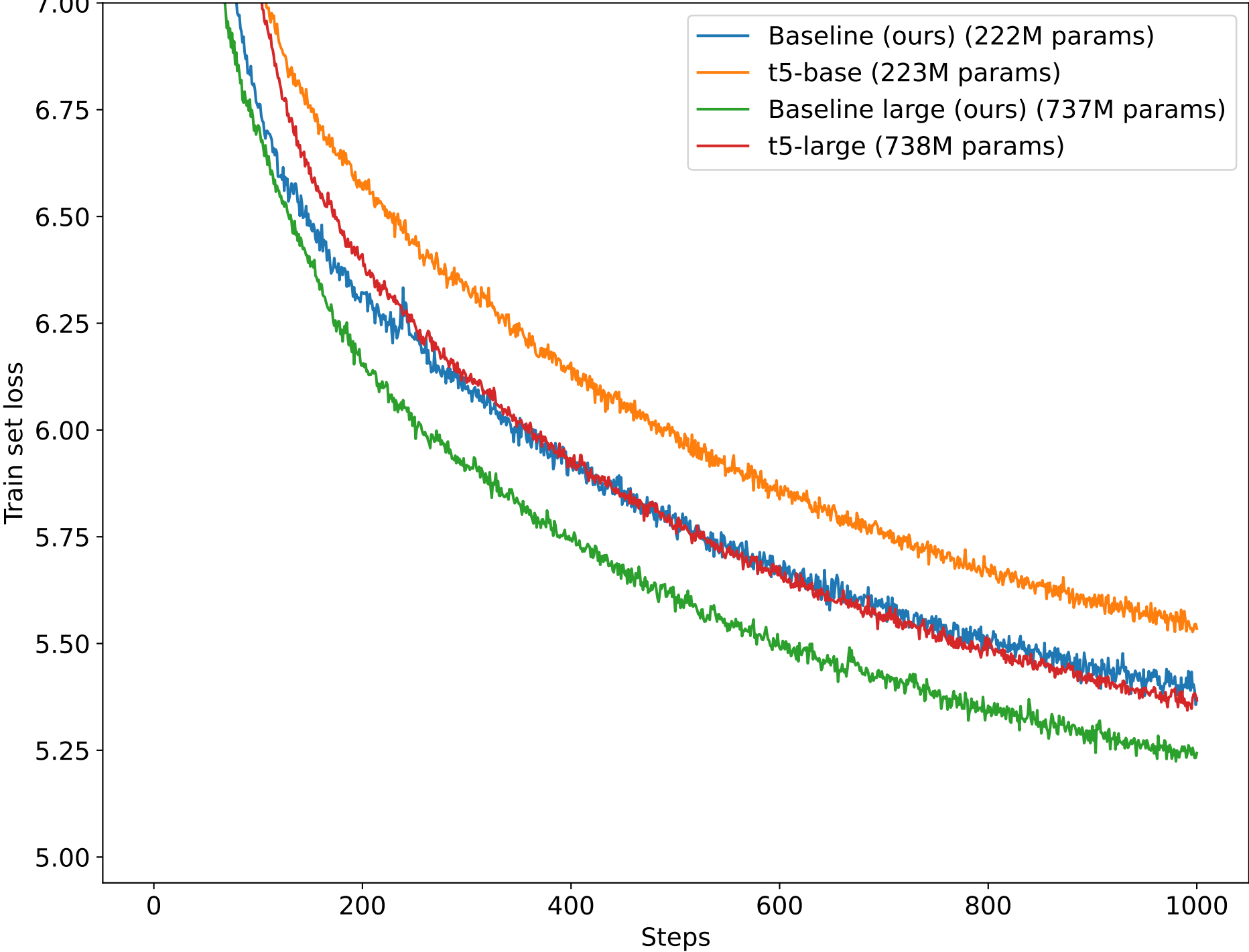
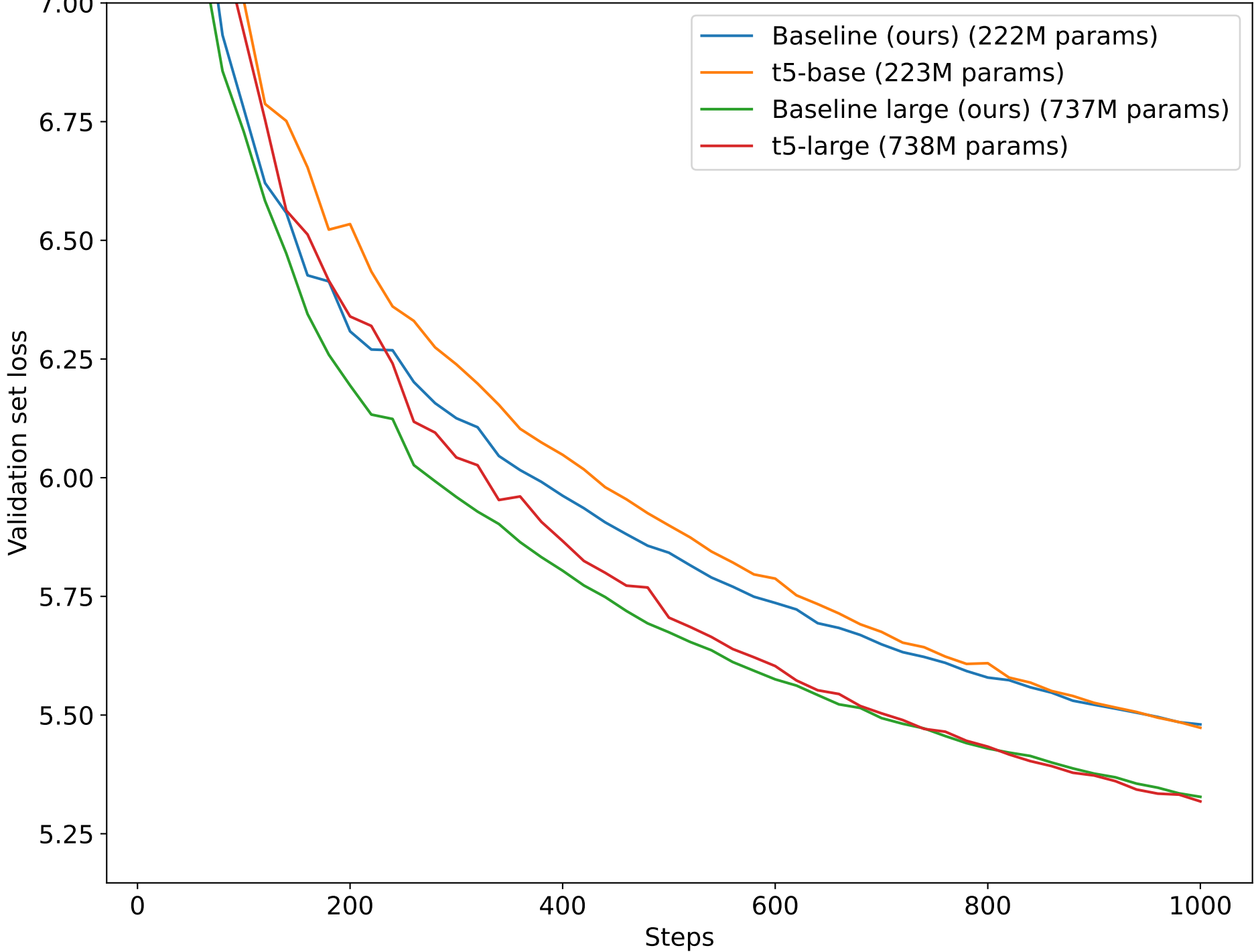


Our reimplementation and t5 baseline

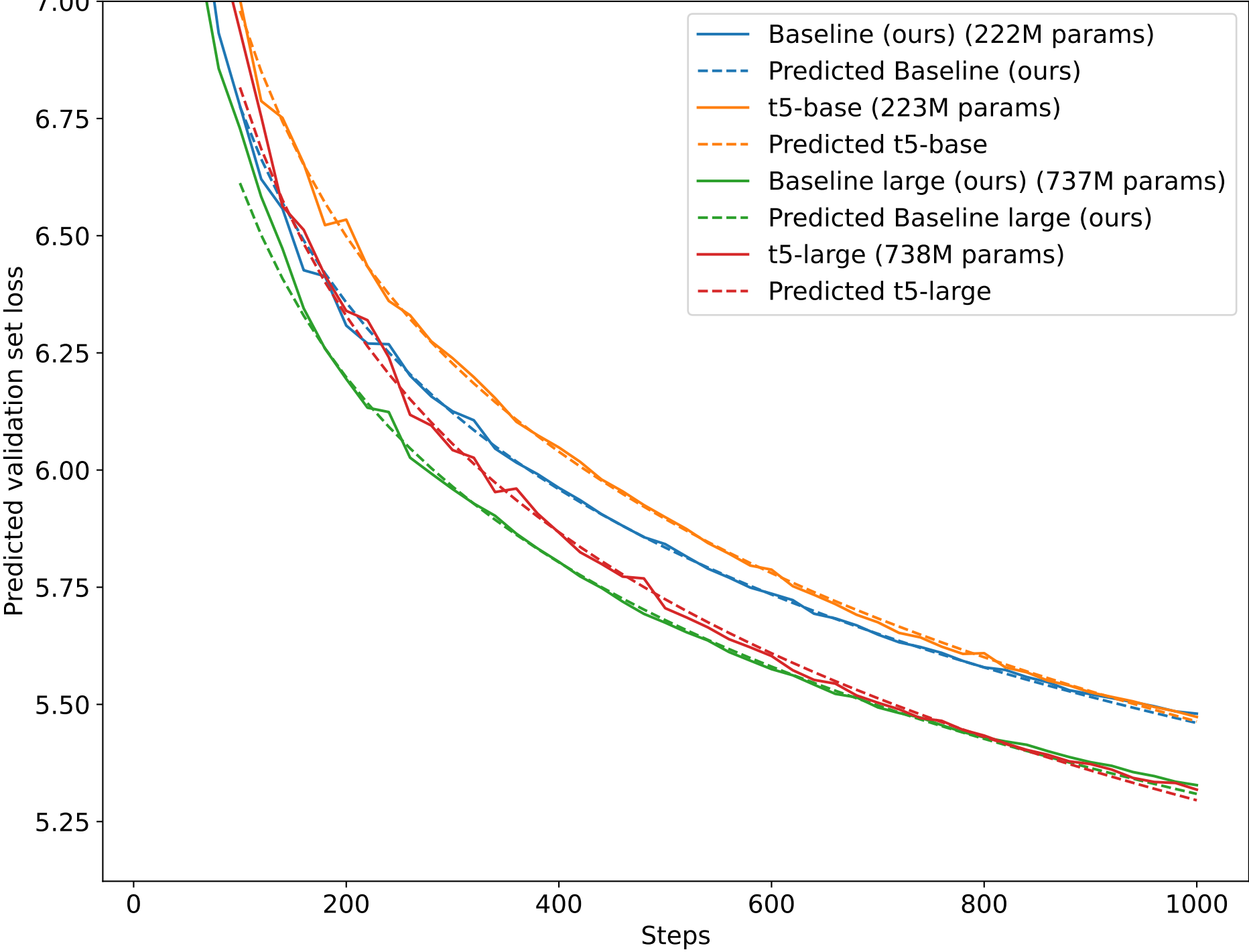
Training loss curves over the 1000 step budget



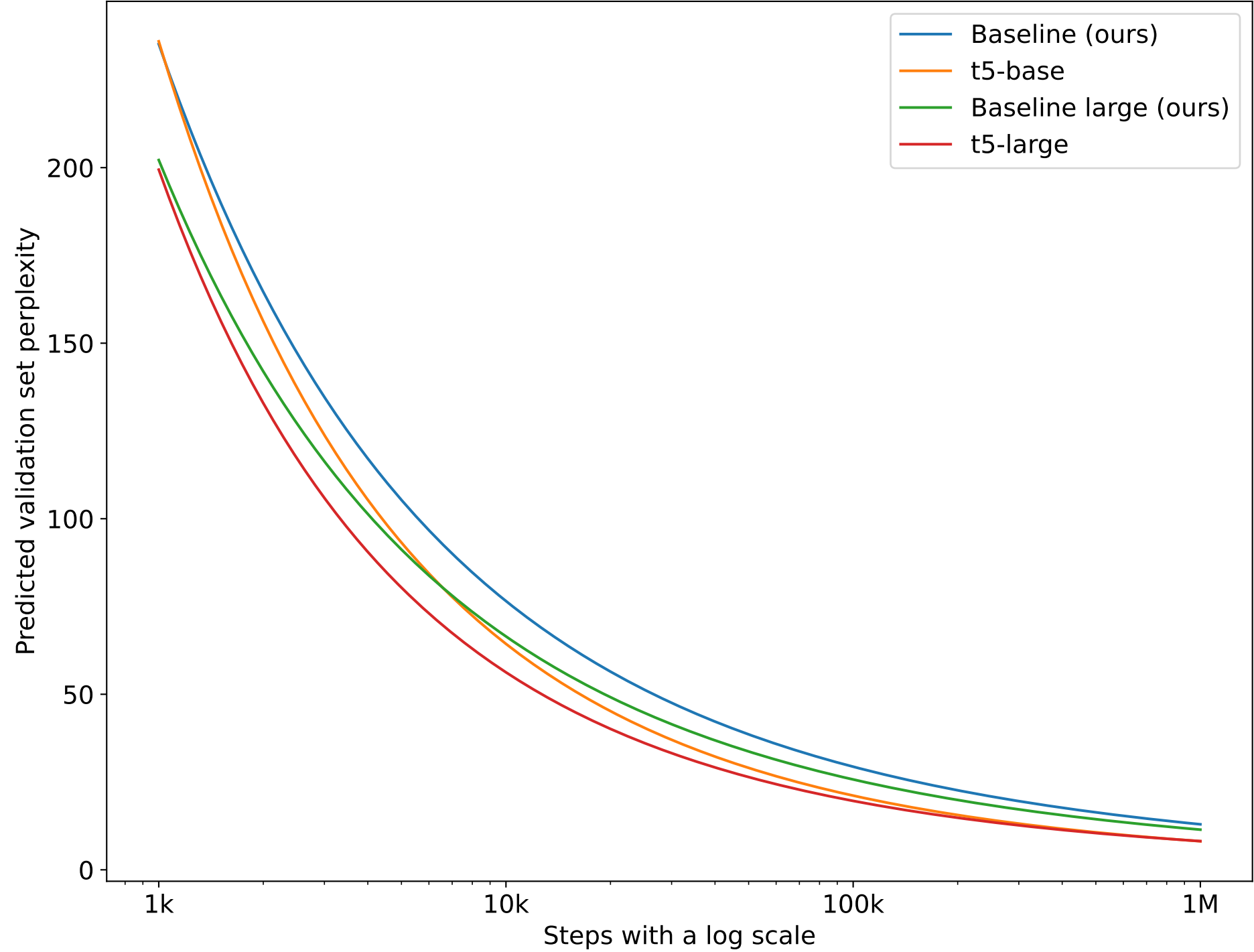
Validation loss curves over the 1000 step budget



Validation loss curves fit to a scaling law



Extrapolated validation perplexity over 1M step budget



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)	$8.98(t^{** -0.069}) - 2.20$	0.009	235.184	76.542	29.363	19.558	12.961
t5-base	$10.61(t^{** -0.067}) - 3.63$	0.009	235.975	64.353	21.129	13.155	8.134
Baseline large (ours)	$8.90(t^{** -0.069}) - 2.29$	0.013	202.160	66.469	25.721	17.196	11.438
t5-large	$9.06(t^{** -0.080}) - 2.24$	0.017	199.426	56.251	19.622	12.659	8.168