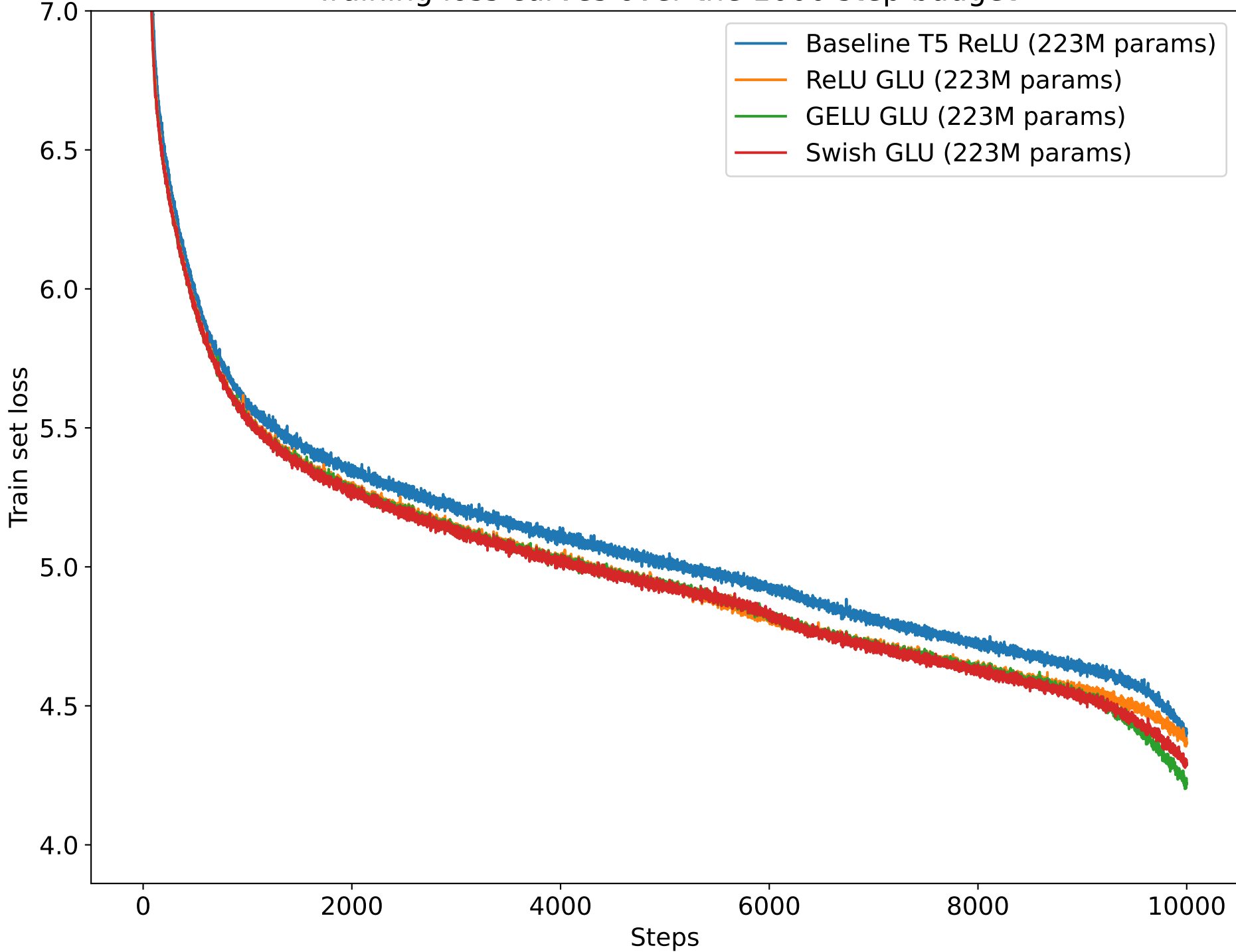
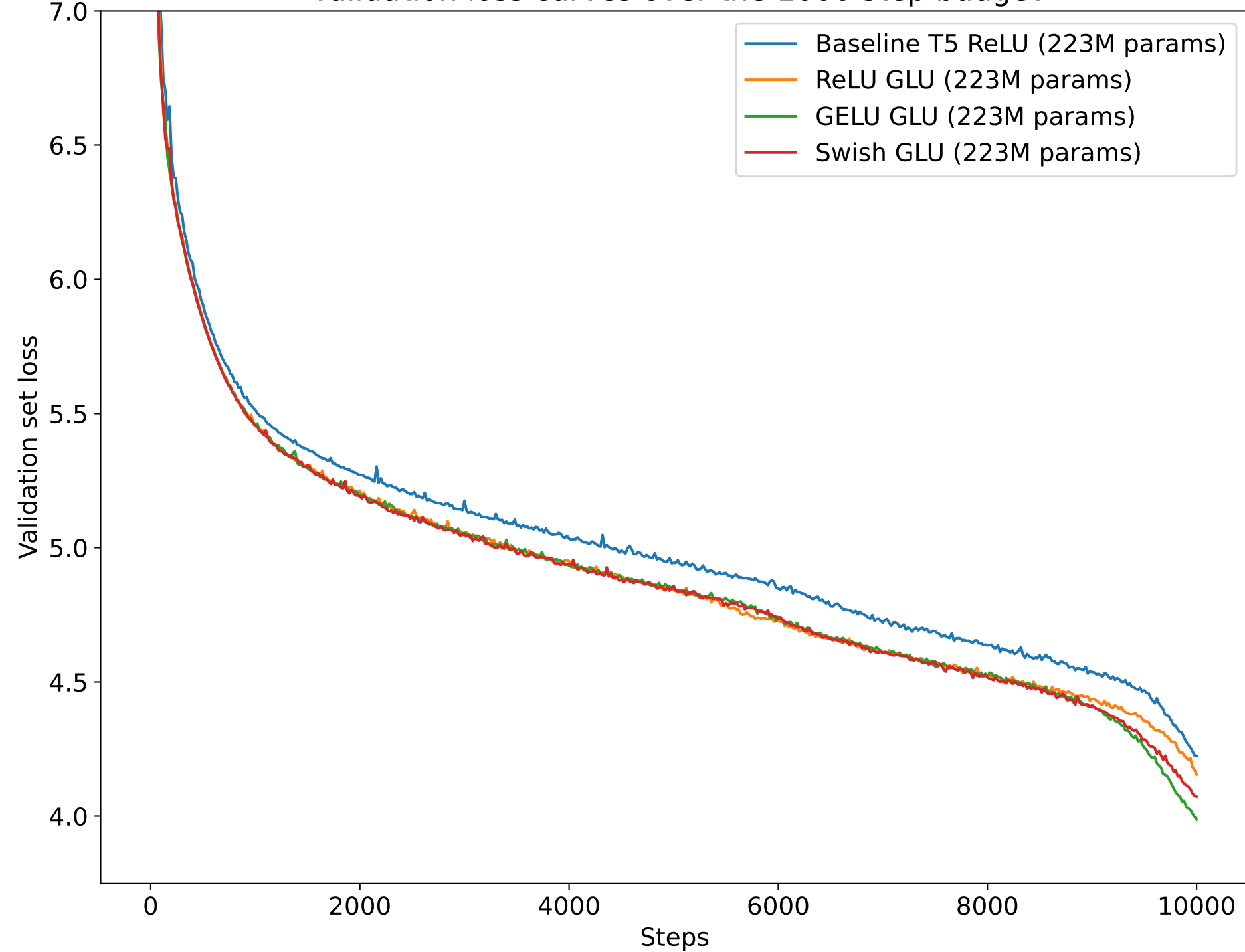


Feedforward layers using Gated Linear Units (GLU)

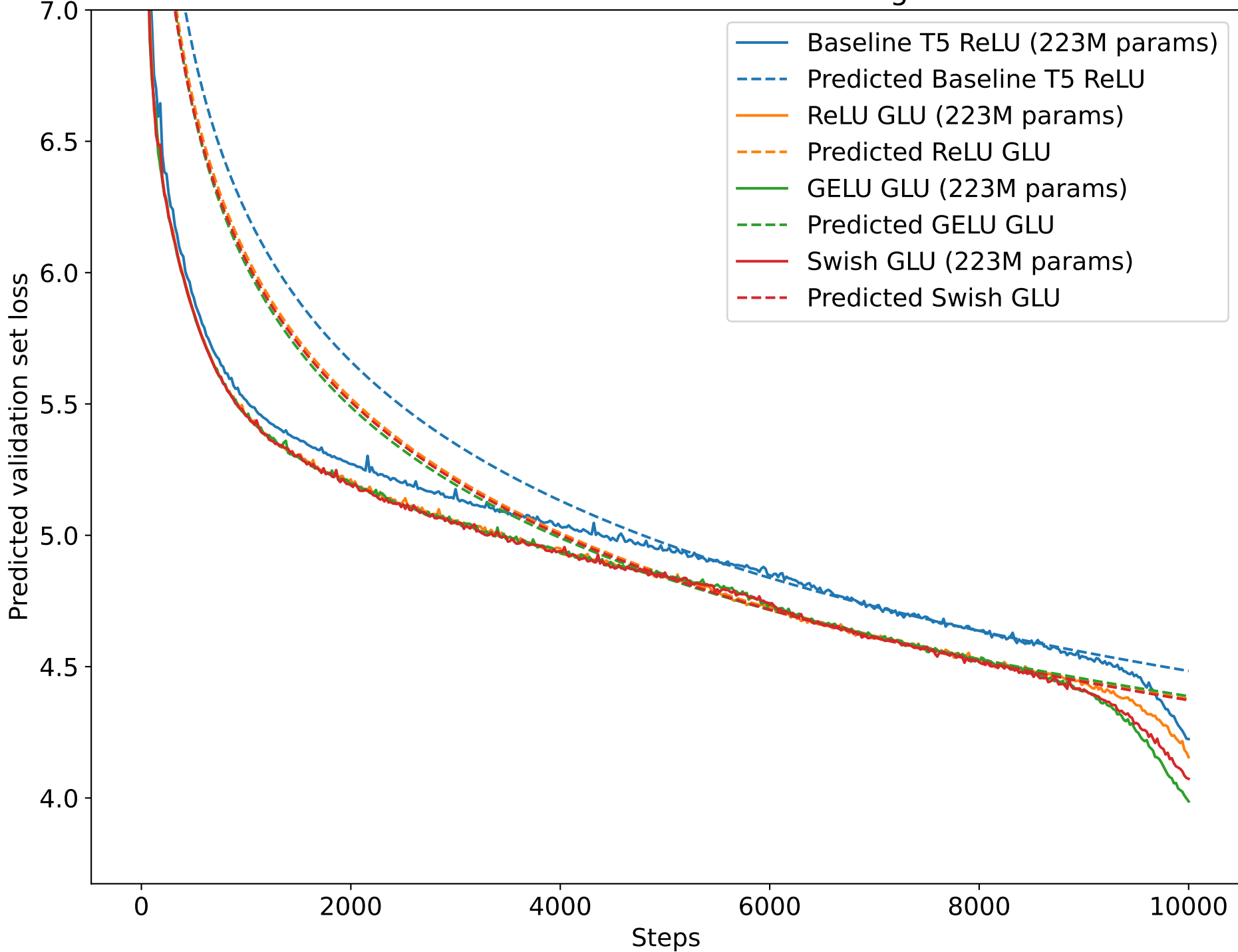
Training loss curves over the 1000 step budget



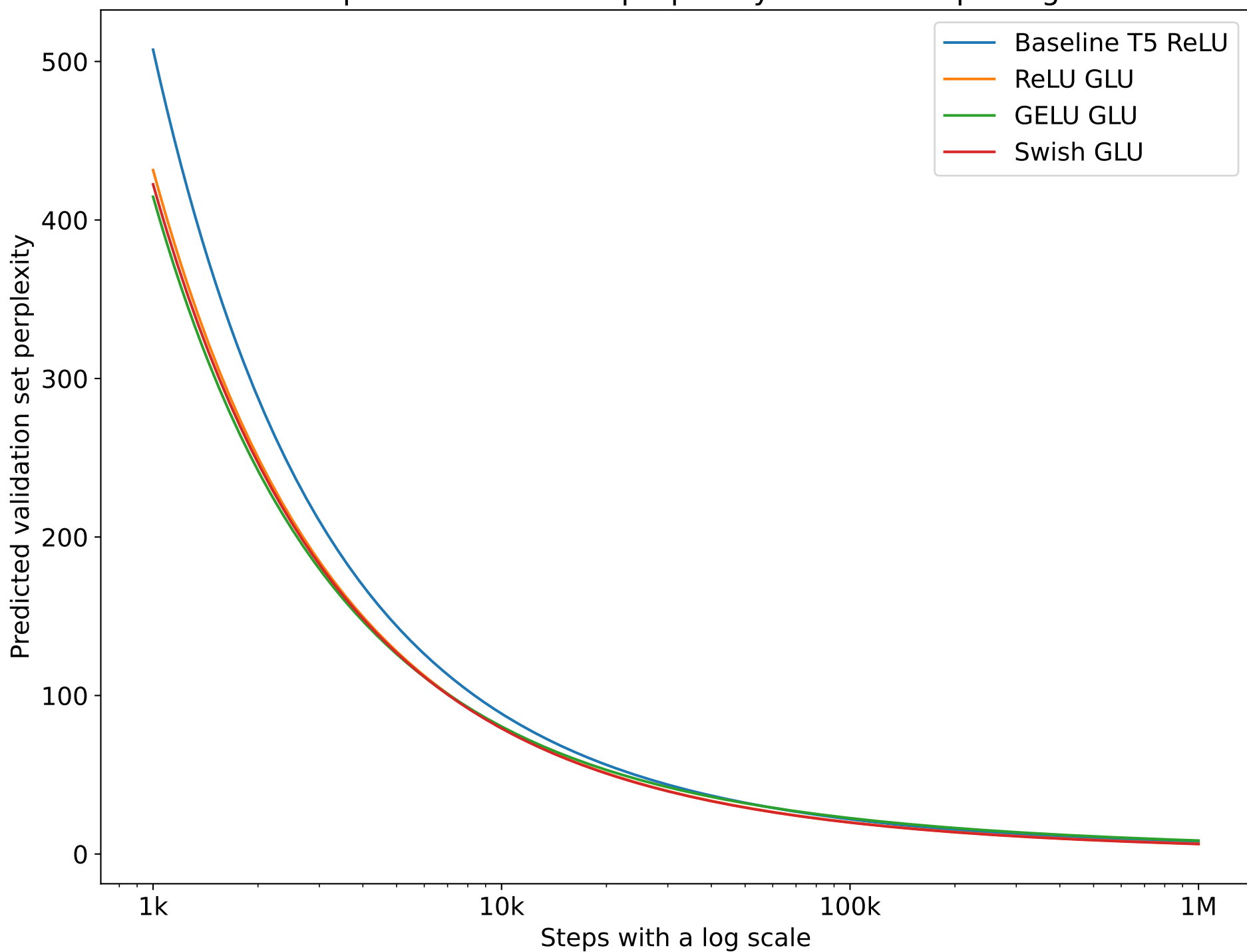
Validation loss curves over the 1000 step budget



Validation loss curves fit to a scaling law



Extrapolated validation perplexity over 1M step budget



Scaling law fit details and perplexity (PPL) predictions

Experiment	Train - eval loss	Scaling law	PPL at 100k	PPL at 300k	PPL at 1M
Baseline T5 ReLU	0.081	$17.04(t^{** -0.097}) - 2.47$	21.931	12.485	7.188
ReLU GLU	0.099	$16.96(t^{** -0.087}) - 3.21$	19.970	11.319	6.448
GELU GLU	0.099	$15.64(t^{** -0.112}) - 1.20$	22.638	13.743	8.496
Swish GLU	0.098	$17.12(t^{** -0.082}) - 3.65$	19.833	11.174	6.307