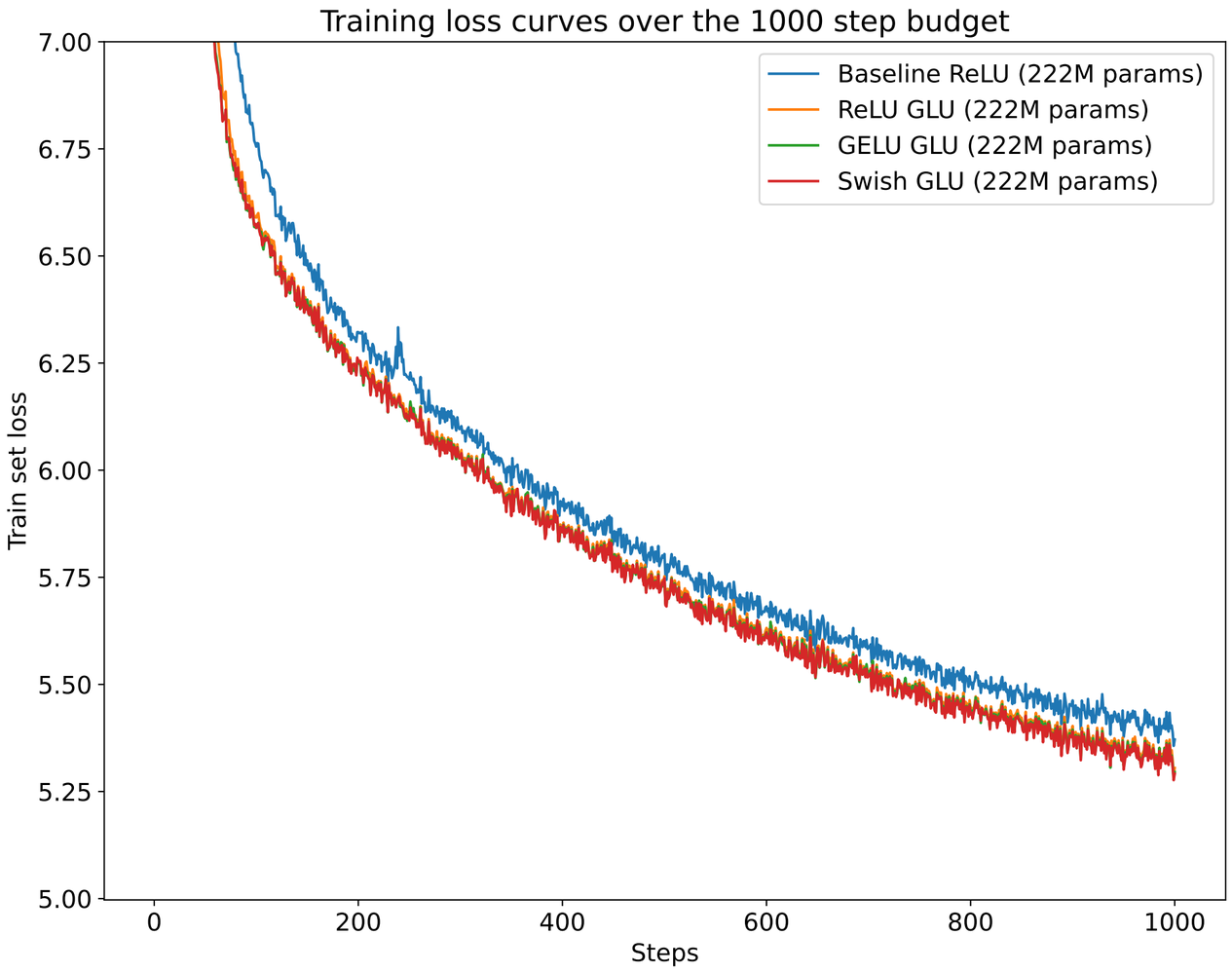
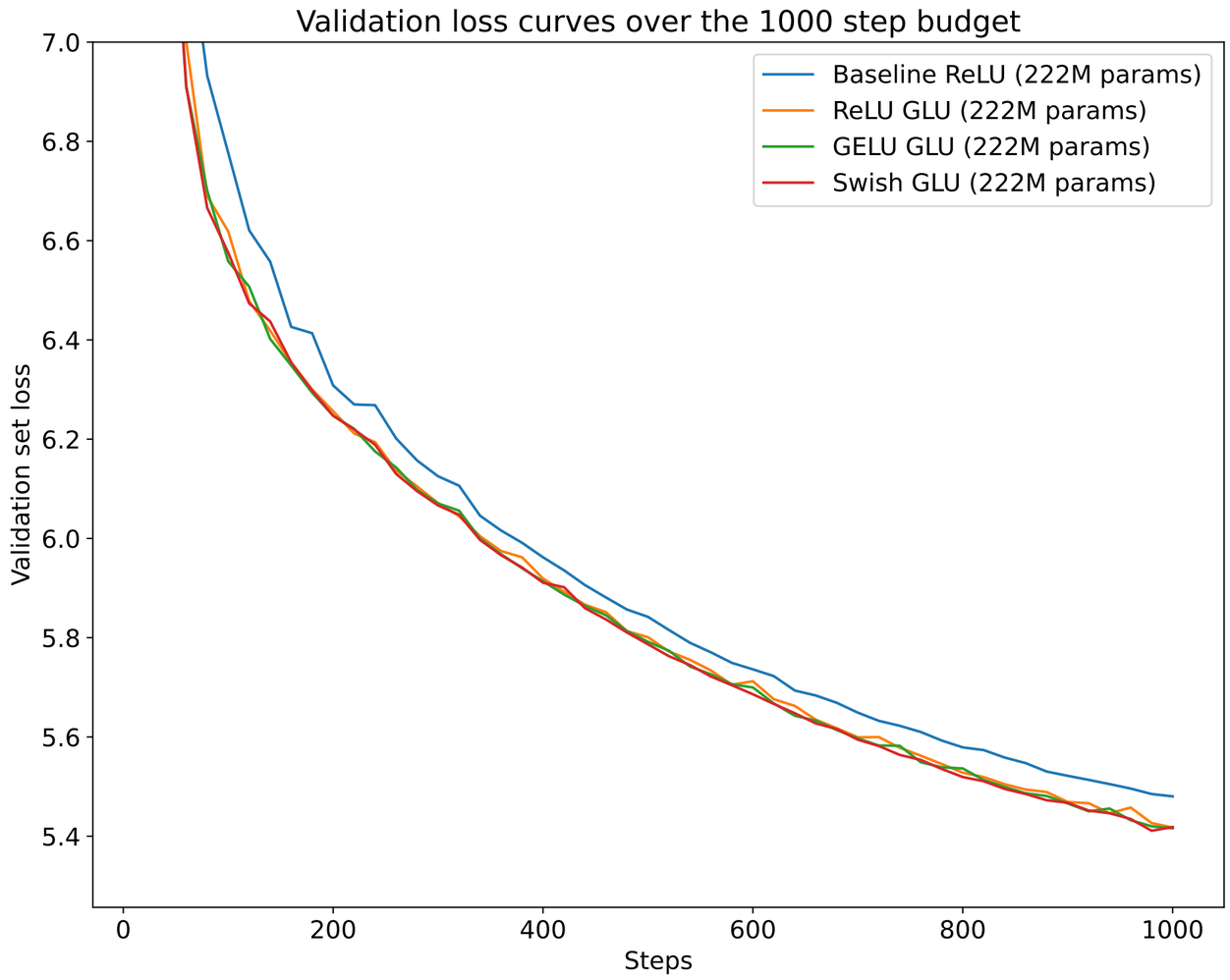
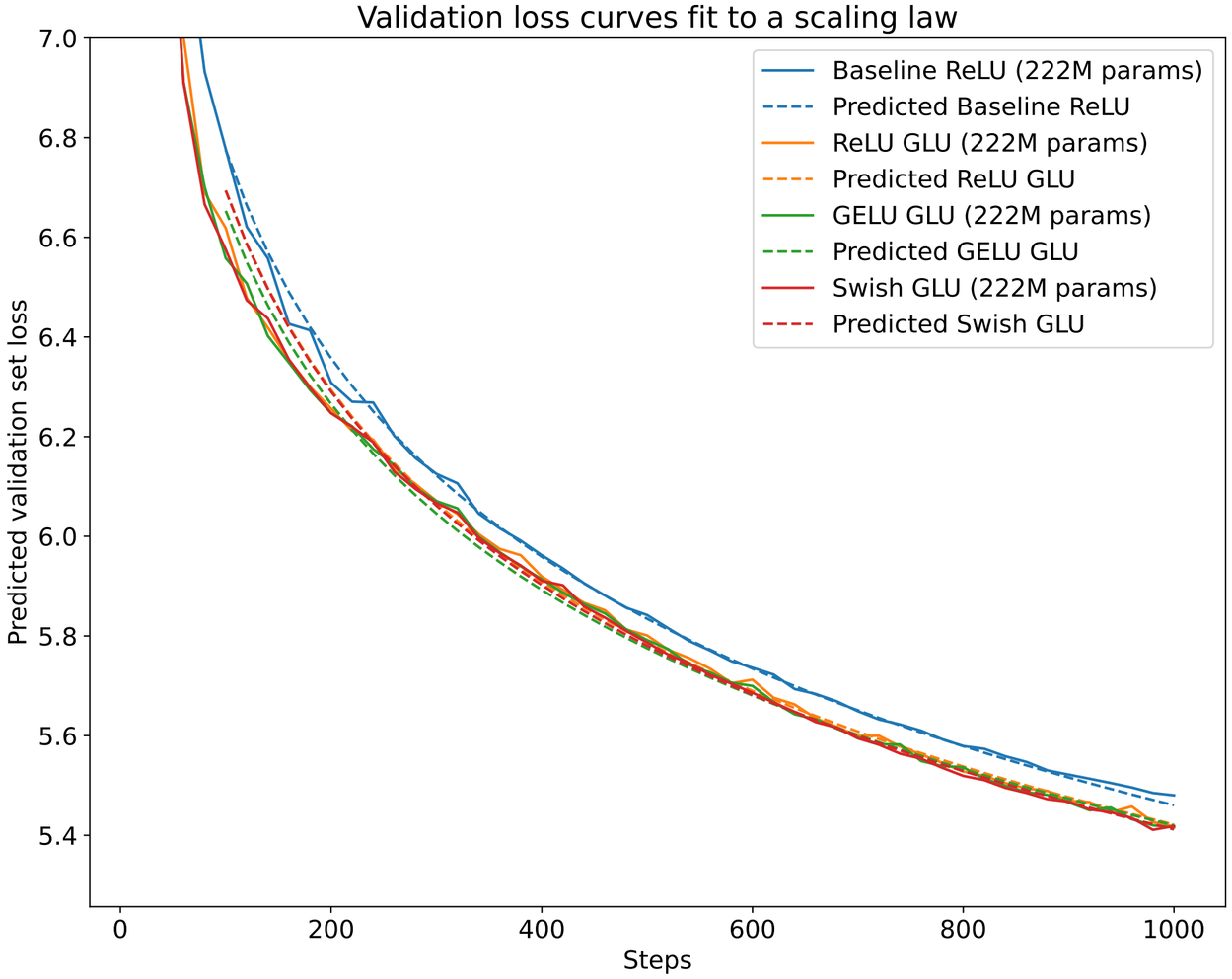
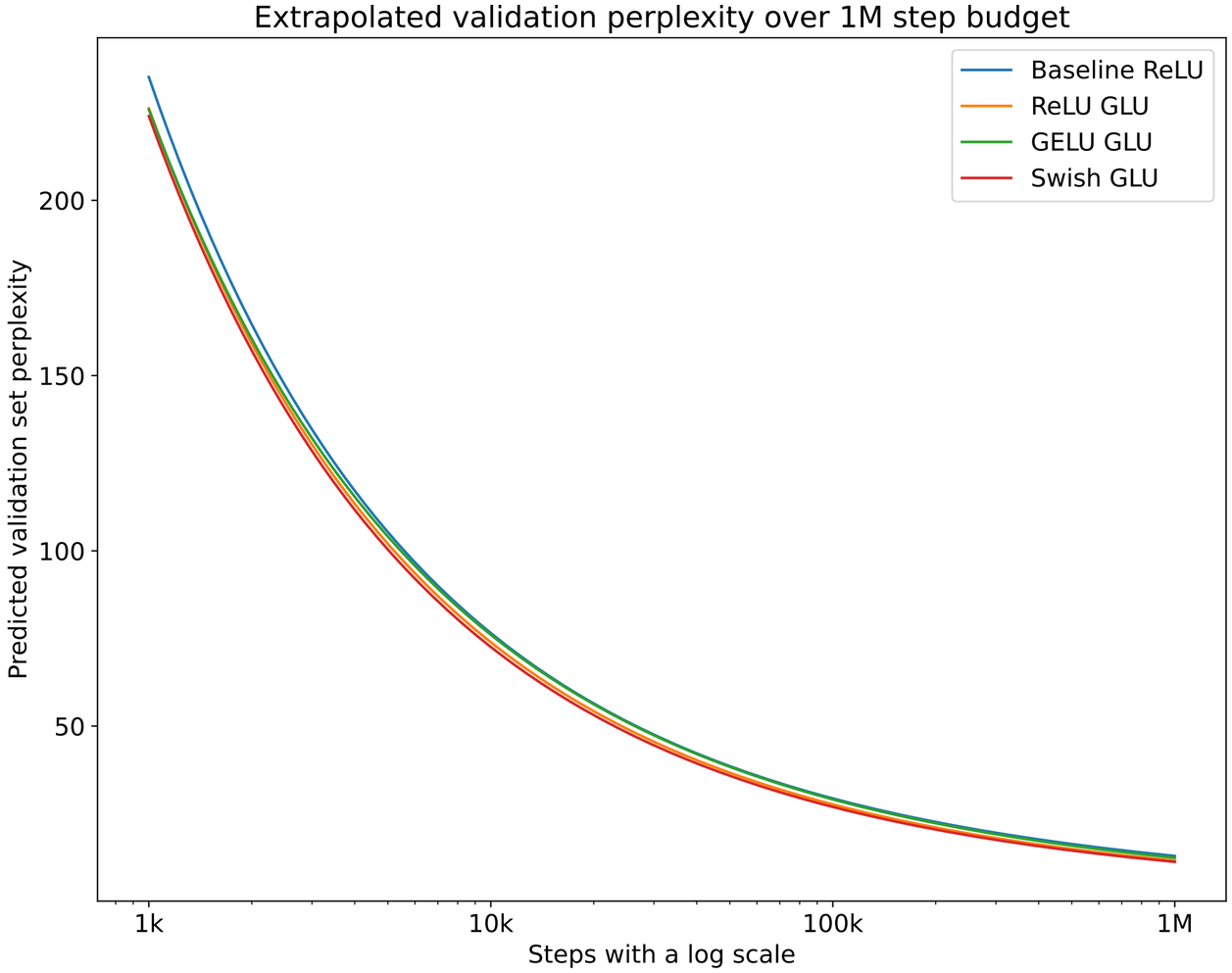
Feedforward layers using Gated Linear Units (GLU)









Experiment	Scaling law	Mean L1 residual	PPL at 1k steps	PPL at 10k steps	PPL at 100k steps	PPL at 300k steps	PPL at 1M steps
Baseline ReLU	8.98(t ** -0.069) - 2.20	0.009	235.184	76.542	29.363	19.558	12.961
ReLU GLU	10.53(t ** -0.056) - 3.84	0.016	226.252	73.930	27.653	18.053	11.648
GELU GLU	10.53(t ** -0.054) - 3.88	0.017	225.896	76.055	29.087	19.148	12.449
Swish GLU	10.61(t ** -0.056) - 3.92	0.016	224.032	72.573	26.939	17.529	11.272