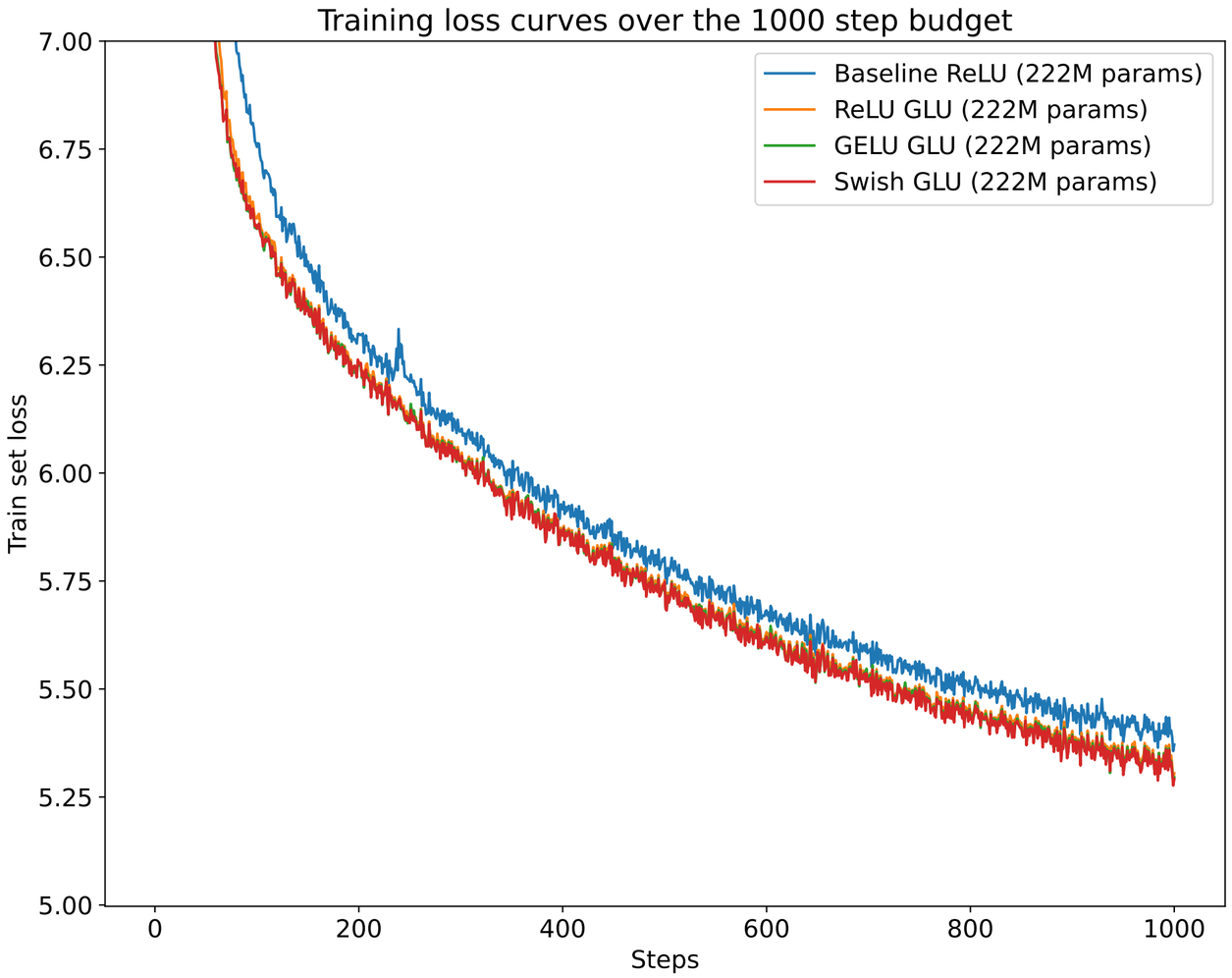
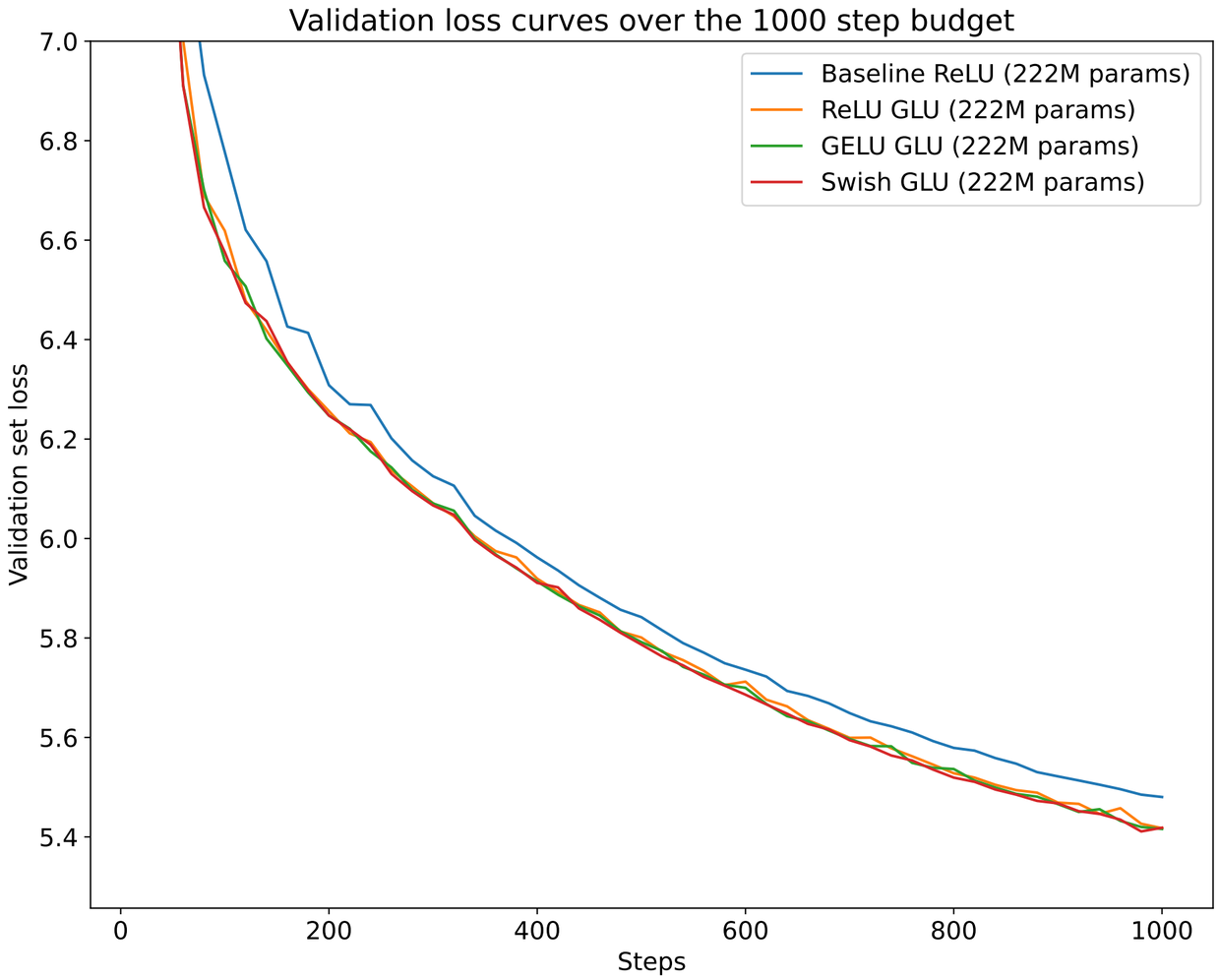
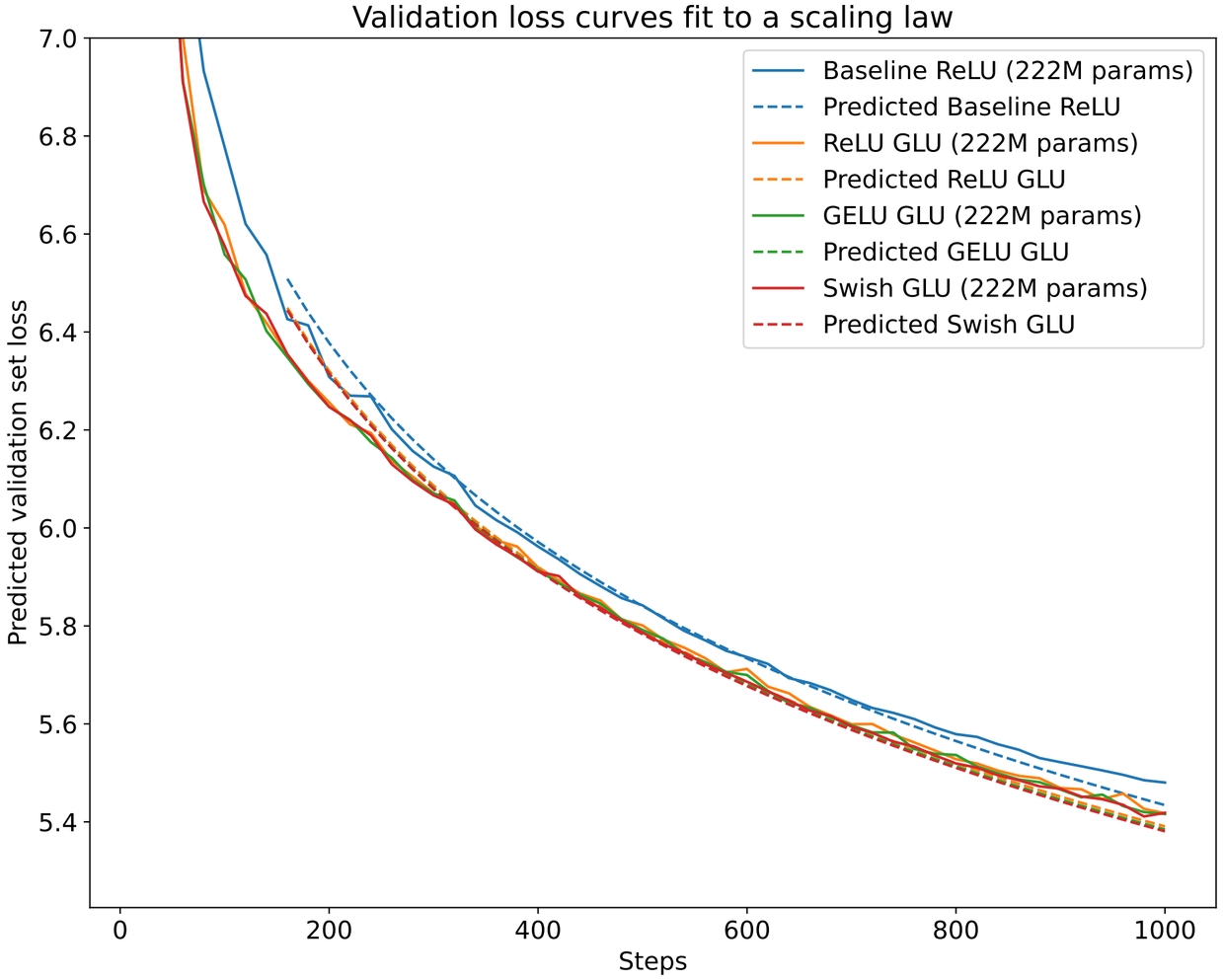
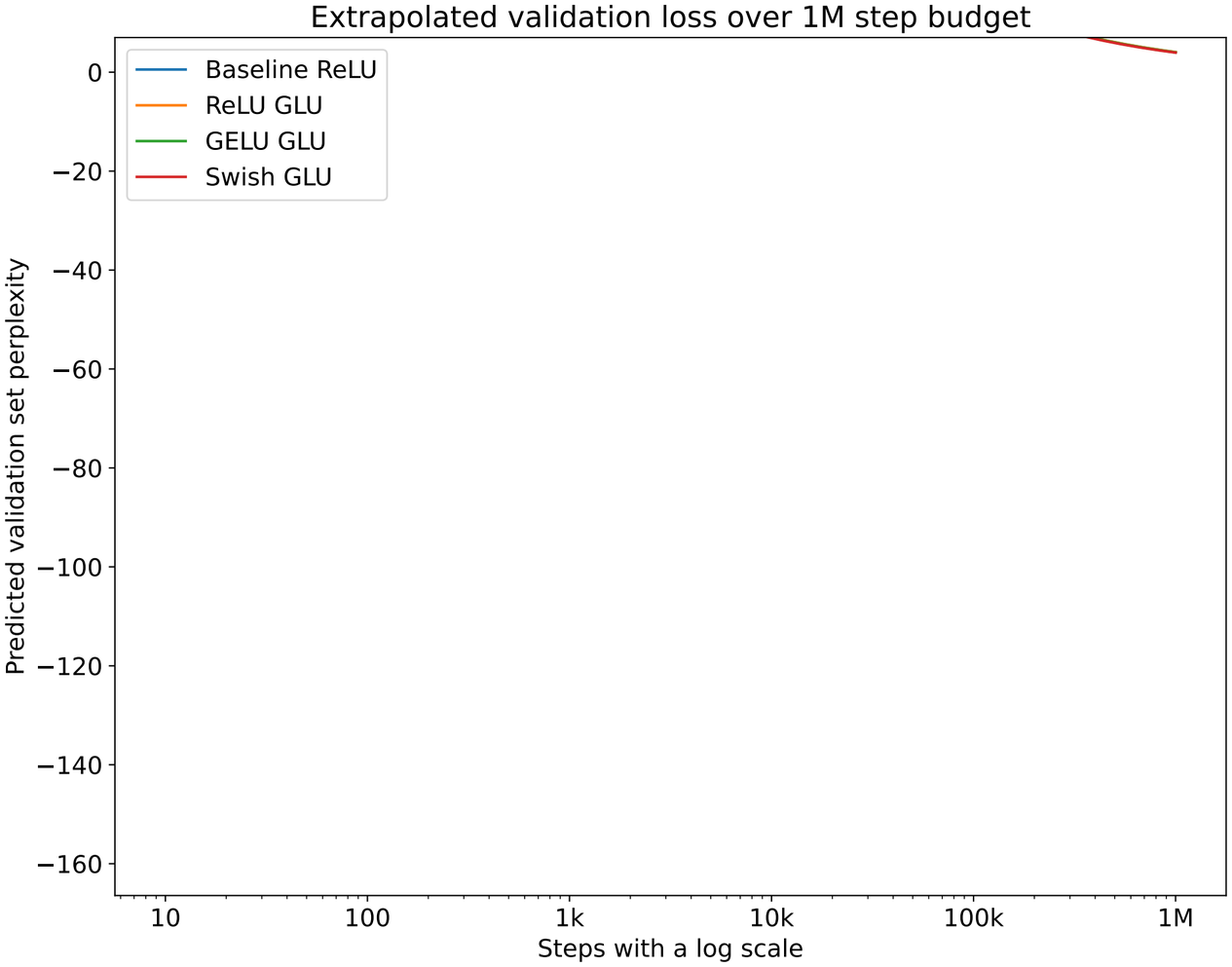
Feedforward layers using Gated Linear Units (GLU)









Experiment	Scaling law	PPL at 1k steps	PPL at 10k steps	PPL at 100k steps	PPL at 300k steps	PPL at 1M steps
Baseline ReLU	2561.15(t ** -0.000) - 2554.37	229.106	59.434	15.429	8.110	4.008
ReLU GLU	8394.56(t ** -0.000) - 8387.84	219.371	58.010	15.344	8.135	4.059
GELU GLU	6943.27(t ** -0.000) - 6936.56	218.021	57.559	15.200	8.053	4.015
Swish GLU	8325.63(t ** -0.000) - 8318.91	217.189	57.085	15.007	7.934	3.946