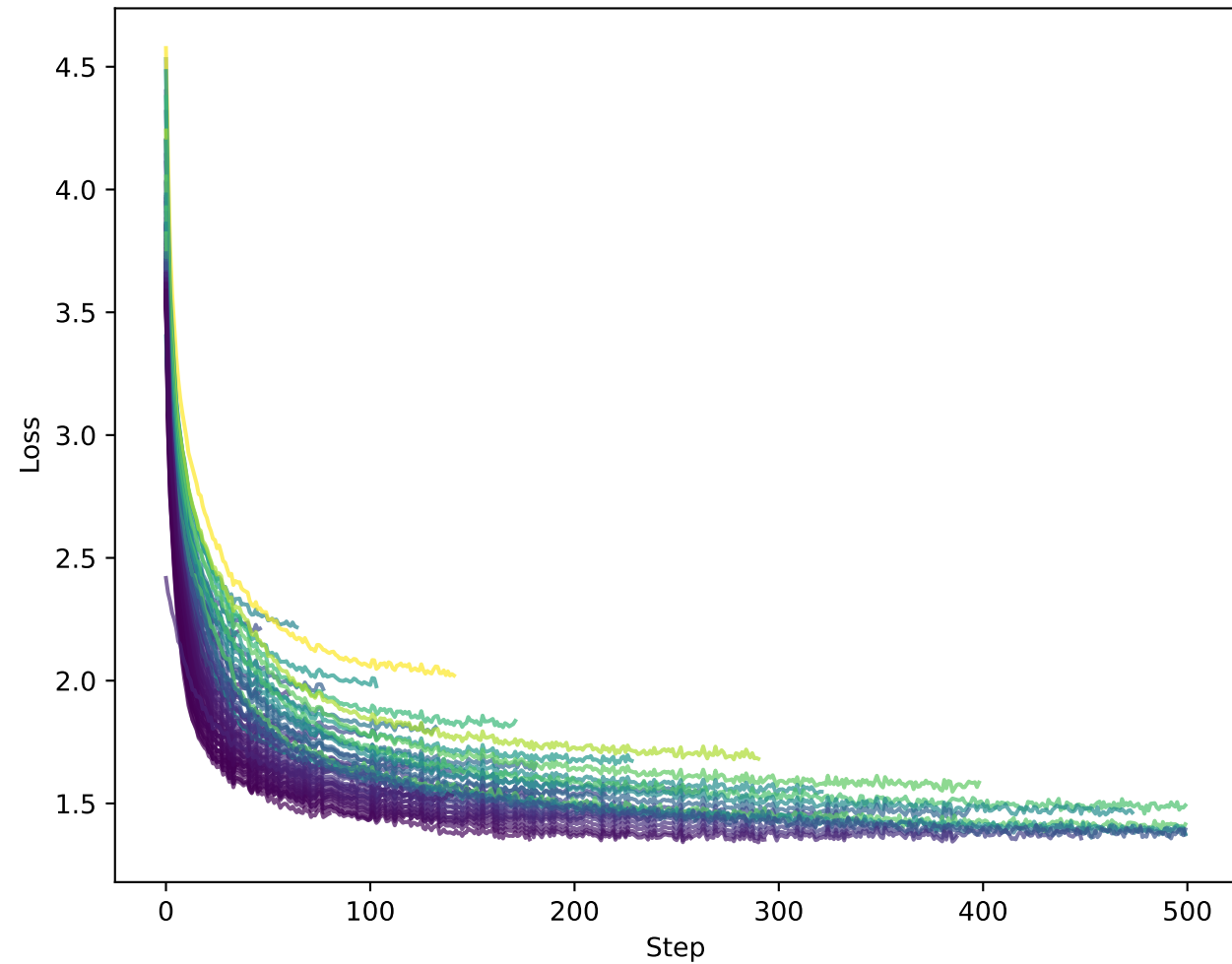bfloat16 weights, bfloat16 activations