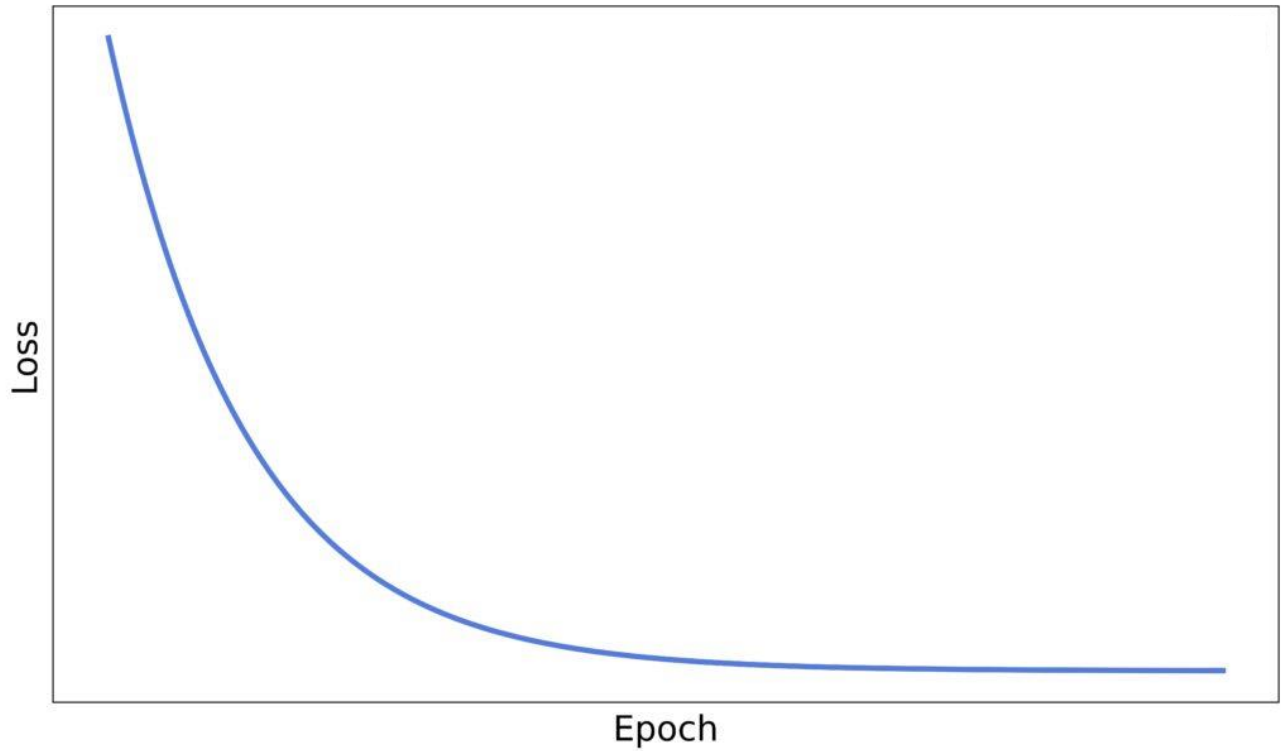
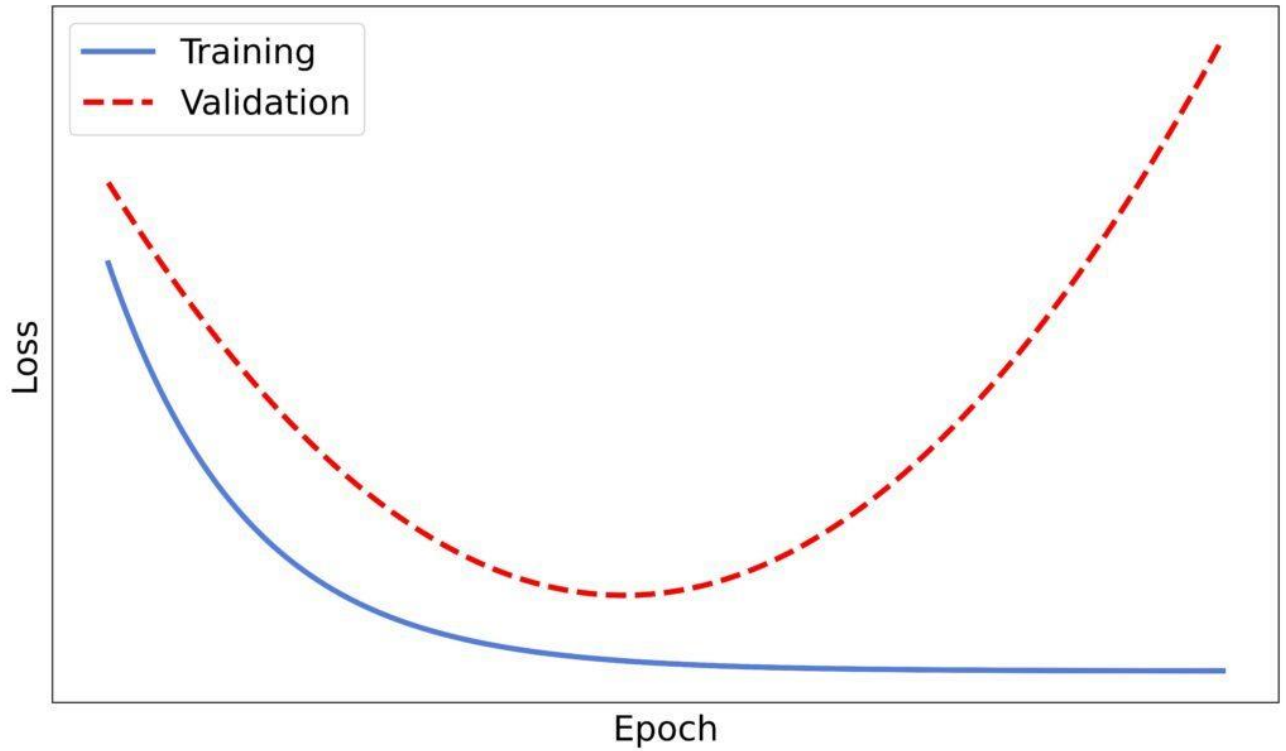
grokking

neural networks: loss graph



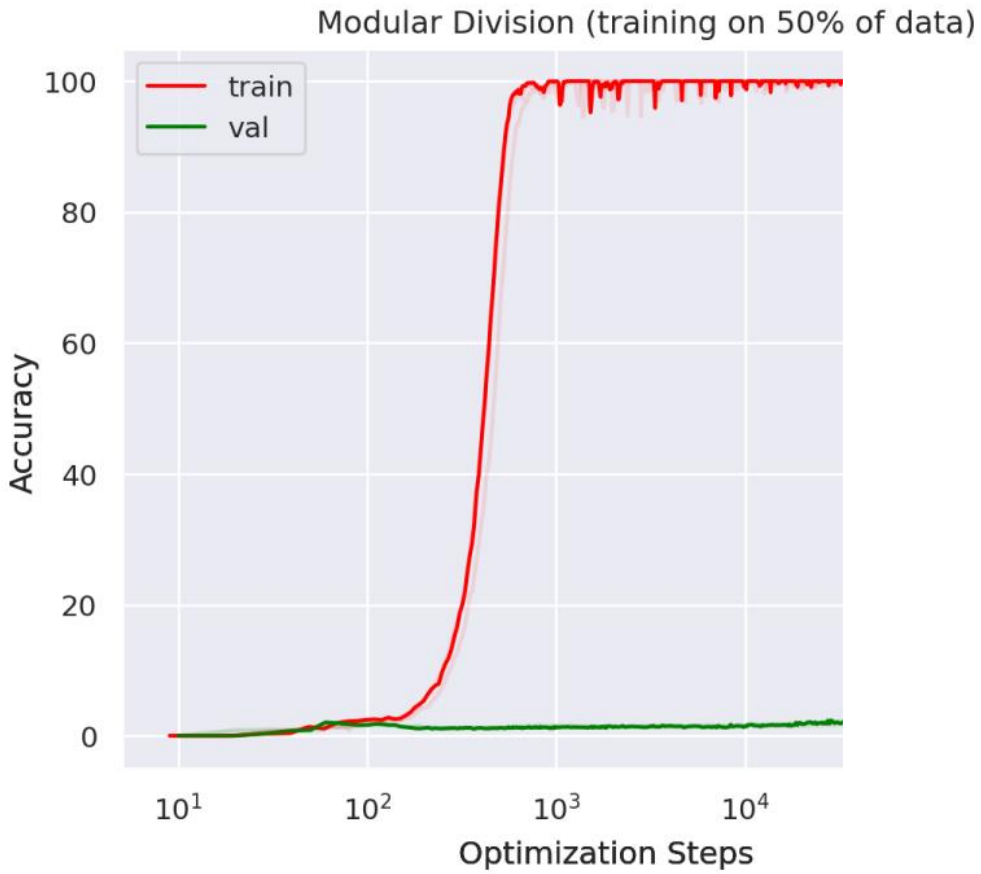
neural networks: overfitting



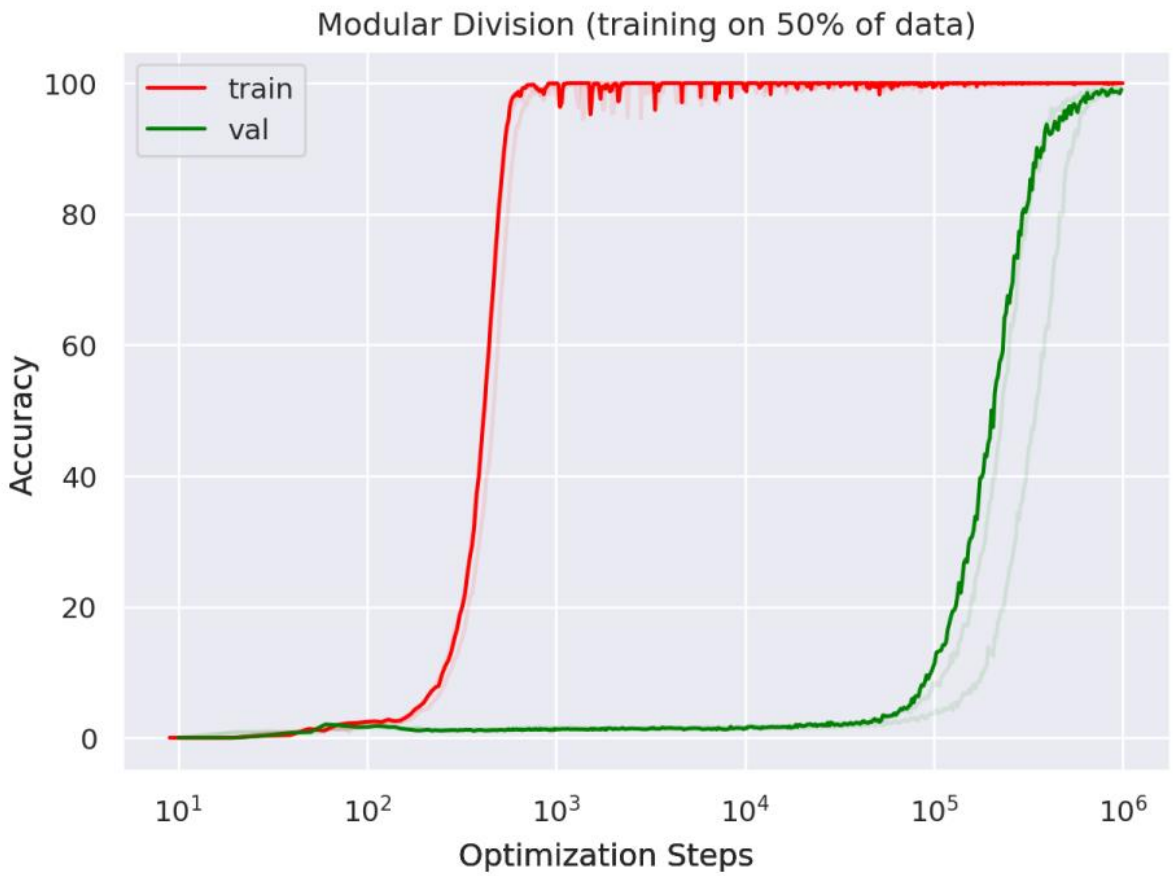
neural networks: weight decay

$$\lambda \|w\|_2$$

grokking: epoch-loss graph

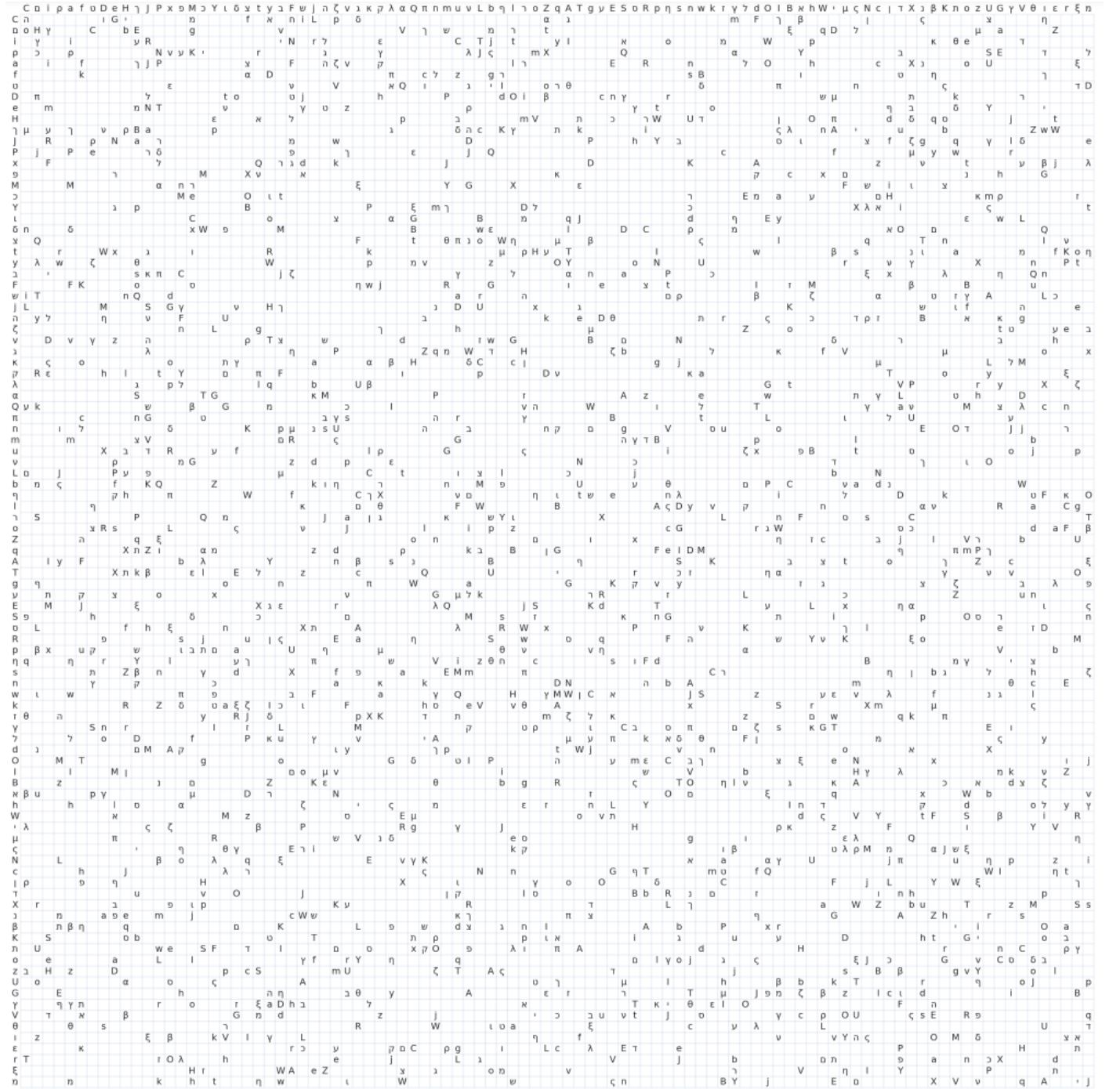


grokking: epoch-loss graph

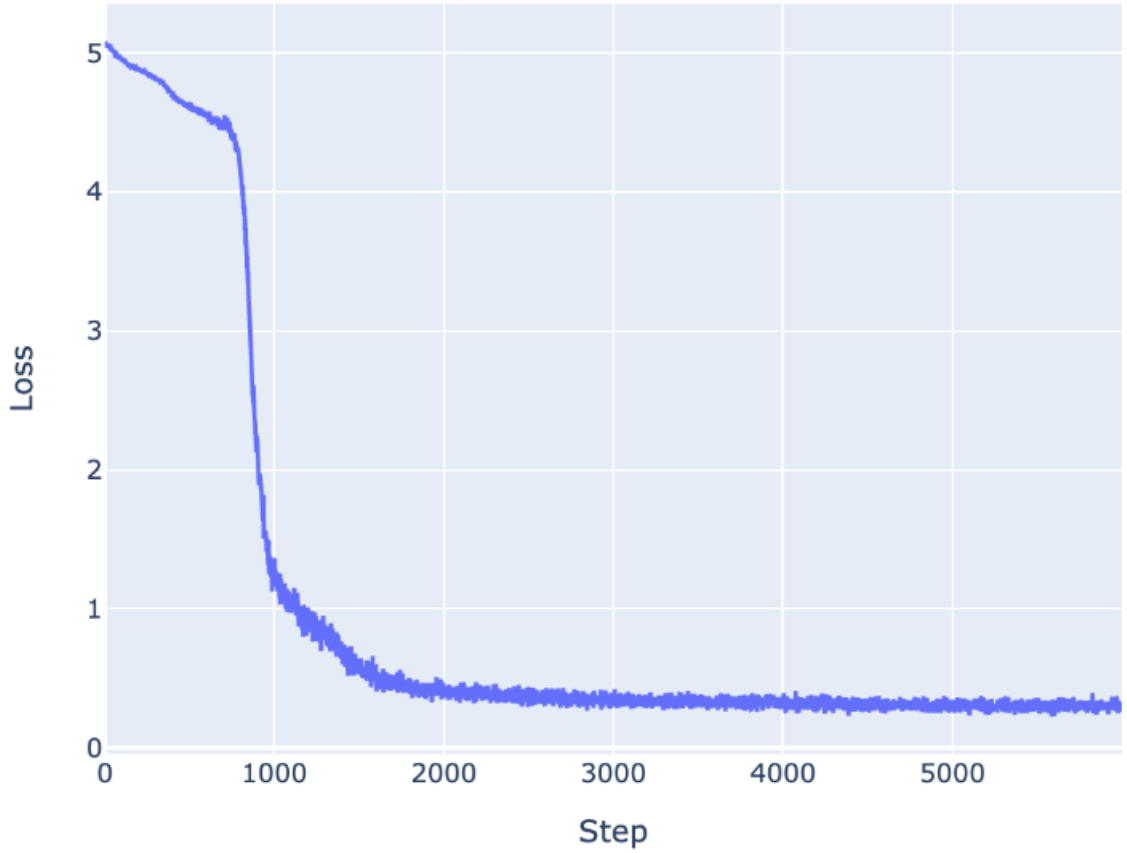


grokking: example

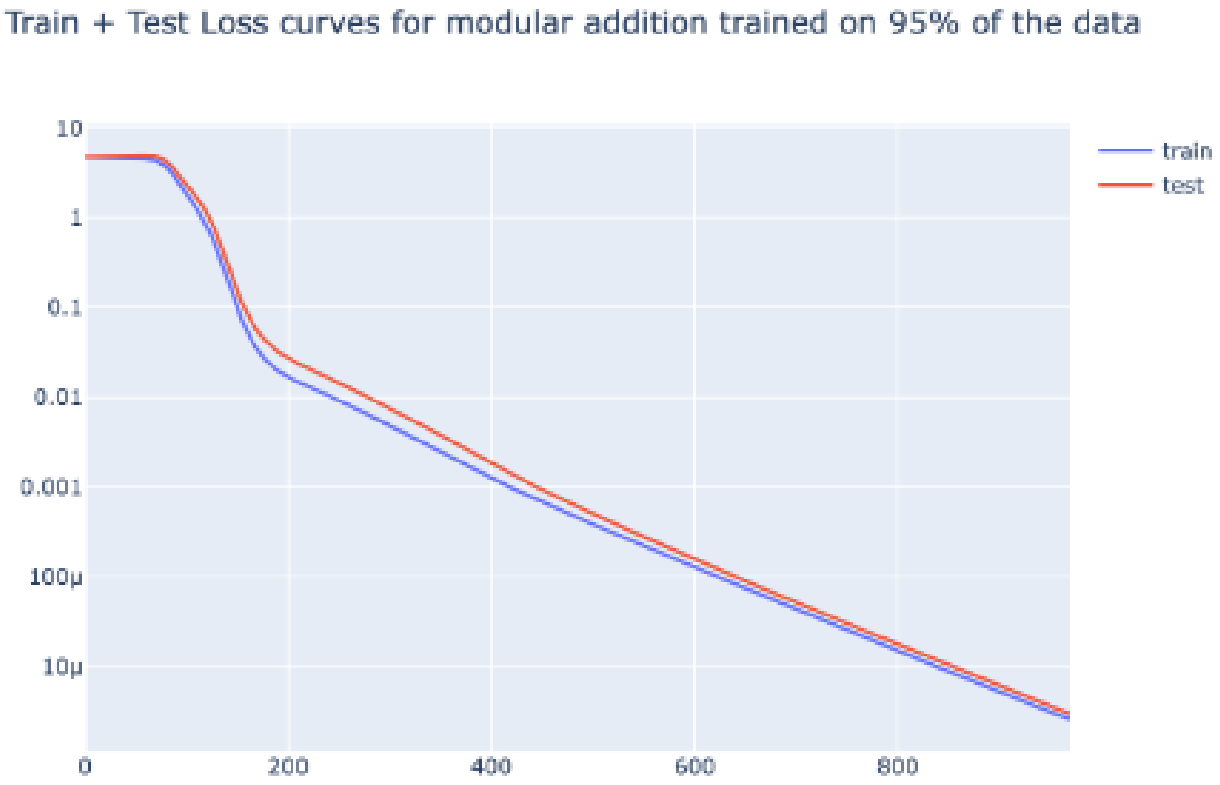
★	a	b	c	d	e
a	a	d	?	c	d
b	c	d	d	a	c
c	?	e	d	b	d
d	a	?	?	b	c
e	b	b	c	?	a



phase change: epoch-loss graph

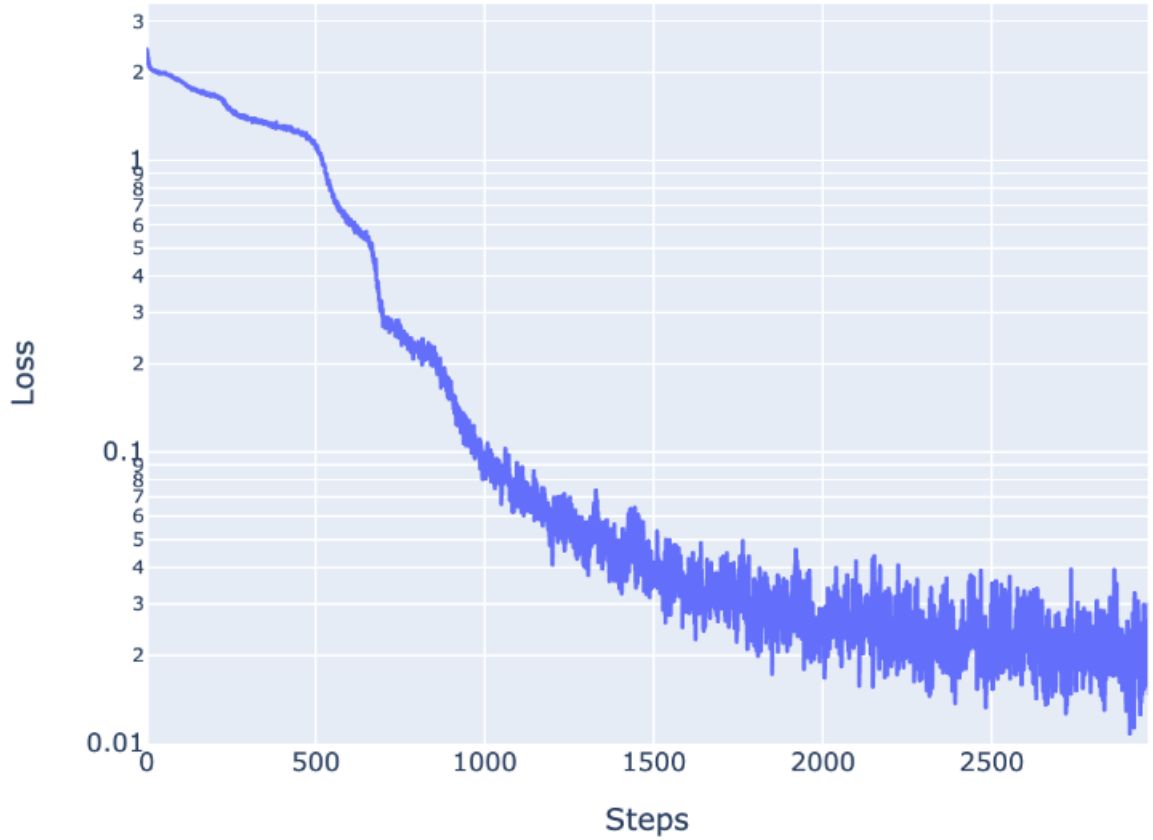


example: addition mod 113

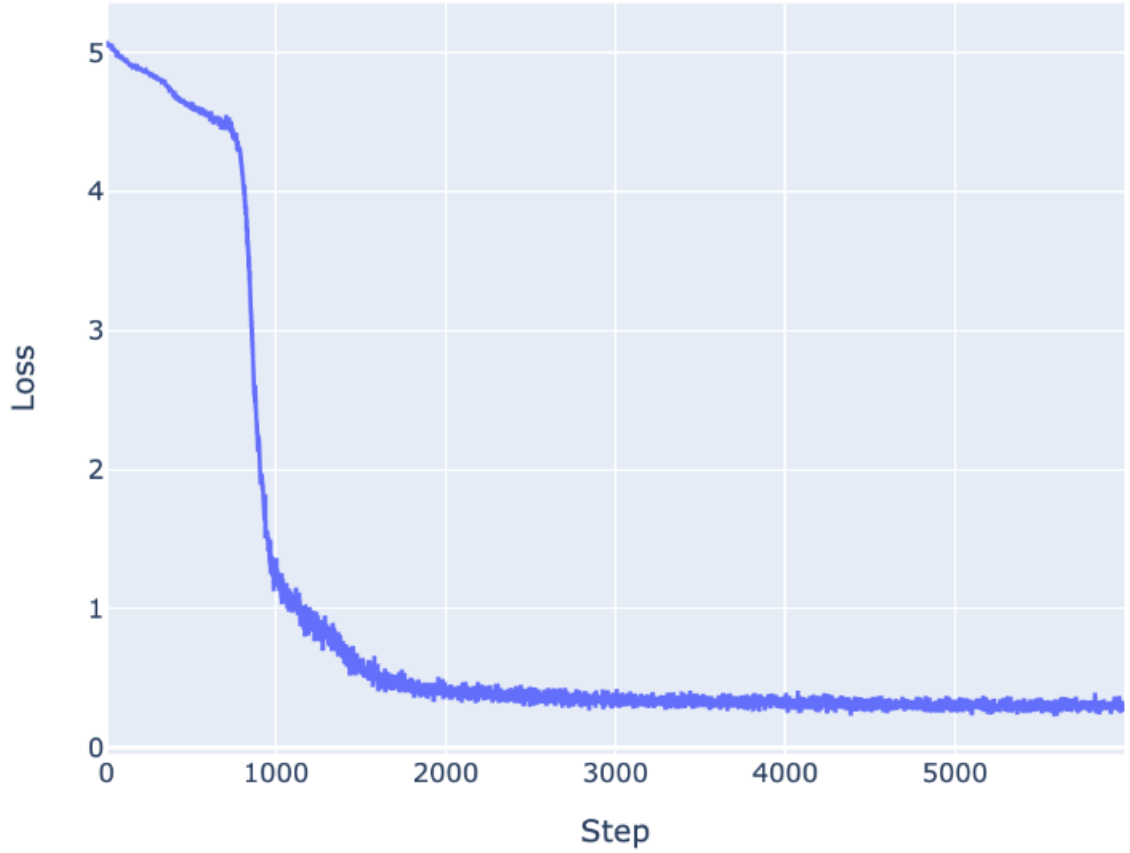


Modular addition mod 113 loss curve, trained on 95% of the data

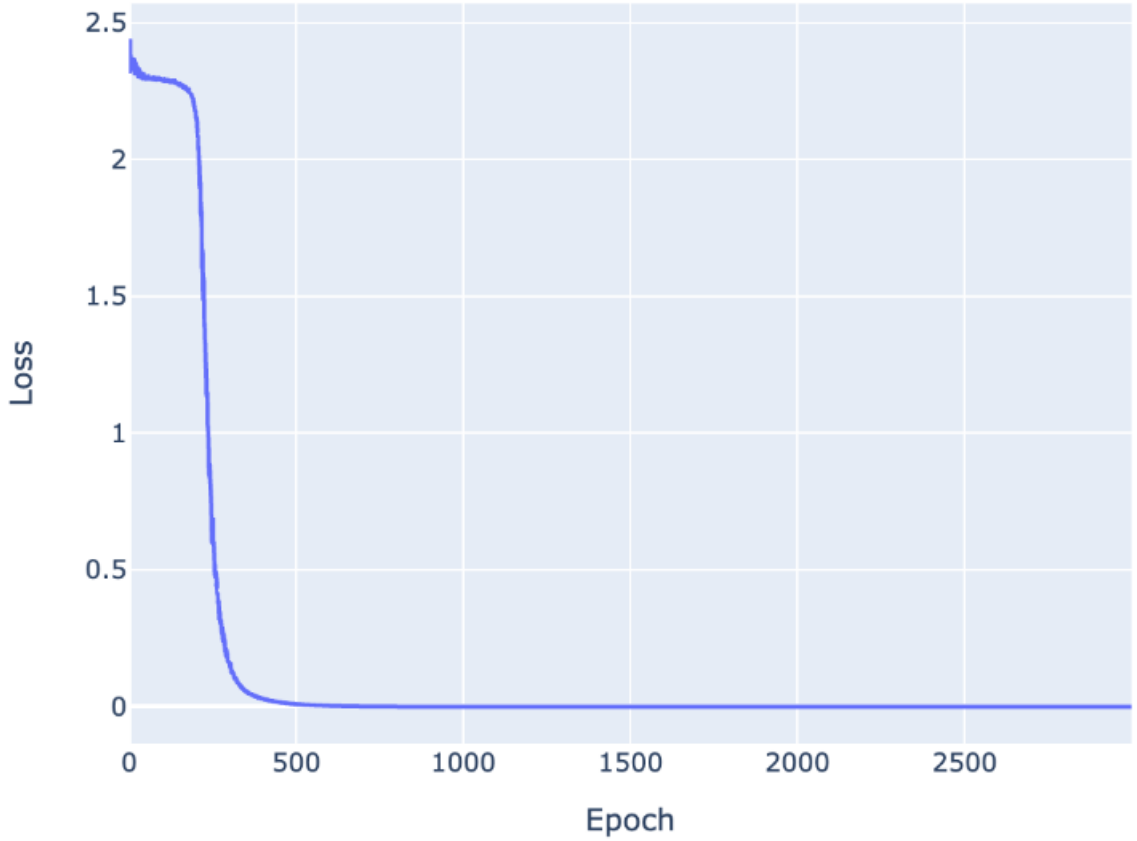
example: 5 digit addition



example: predicting repeated subsequences



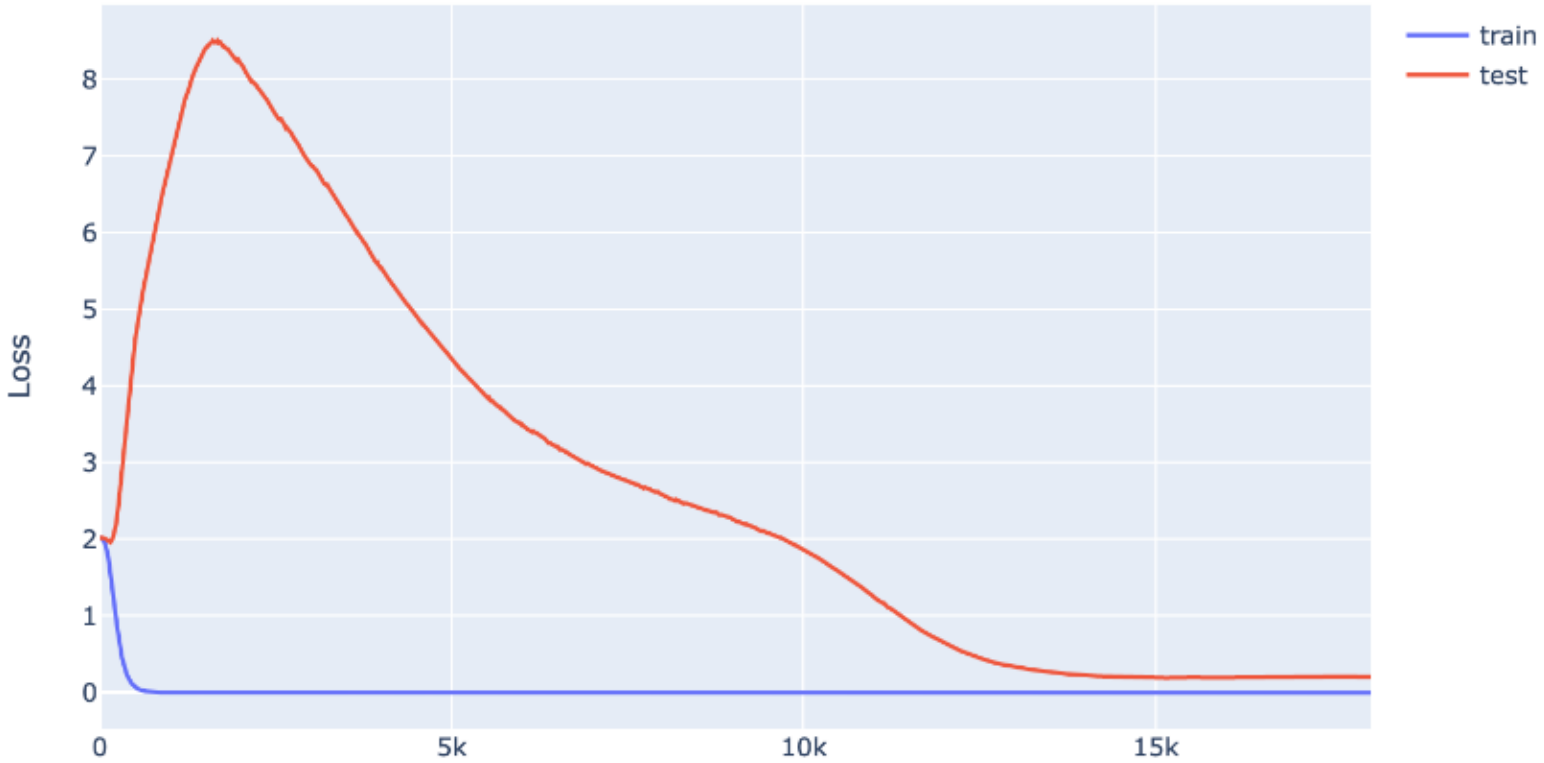
example: finding the max element



phase change: possible explanation

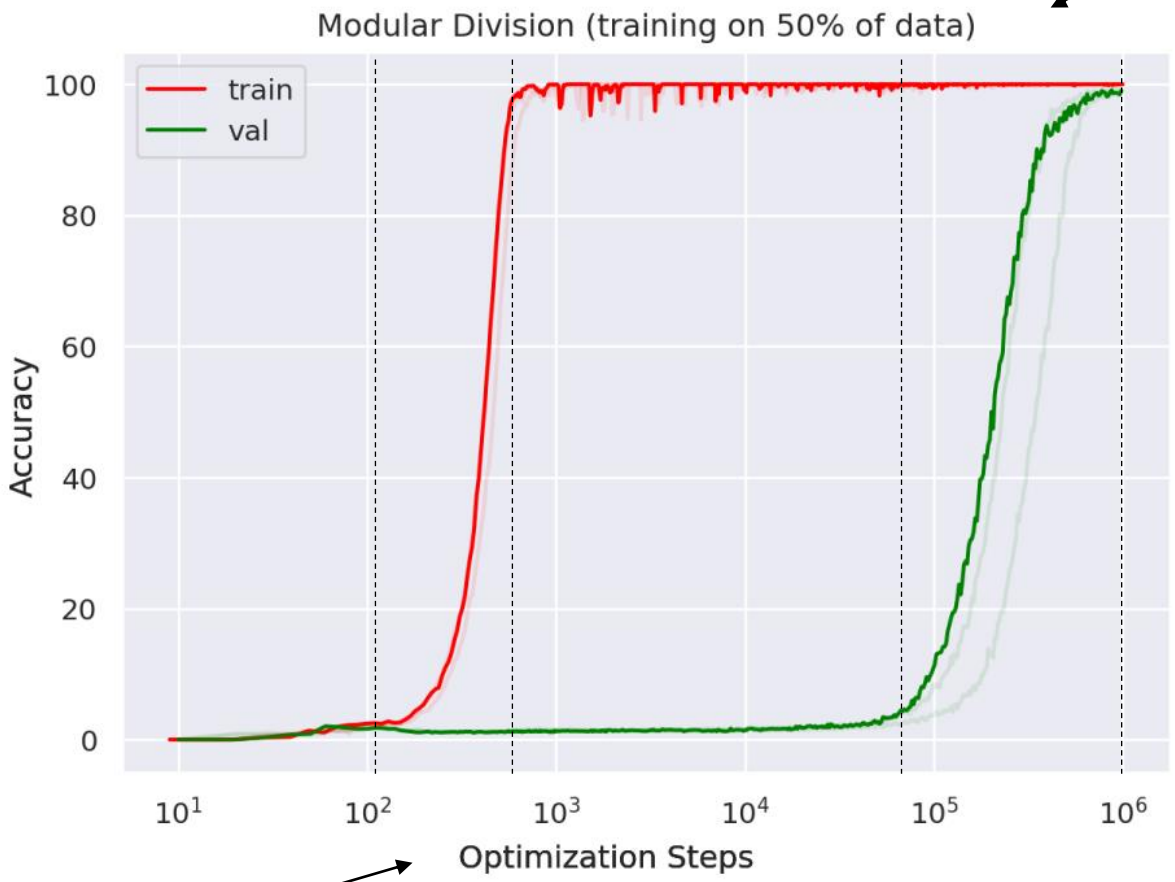
- A lottery ticket hypothesis-inspired explanation
- A random walk explanation
- An evolutionary explanation

phase change + weight decay + small sample



grokking: explanation

phase change



memorization

bonus: $x + y$ mod p algorithm

1. x, y
2. cos kx, sin kx, cos ky, sin ky
3. cos $k(x + y)$, sin $k(x + y)$
4. cos $k(x + y - z)$

$$k = \frac{2\pi}{p}$$