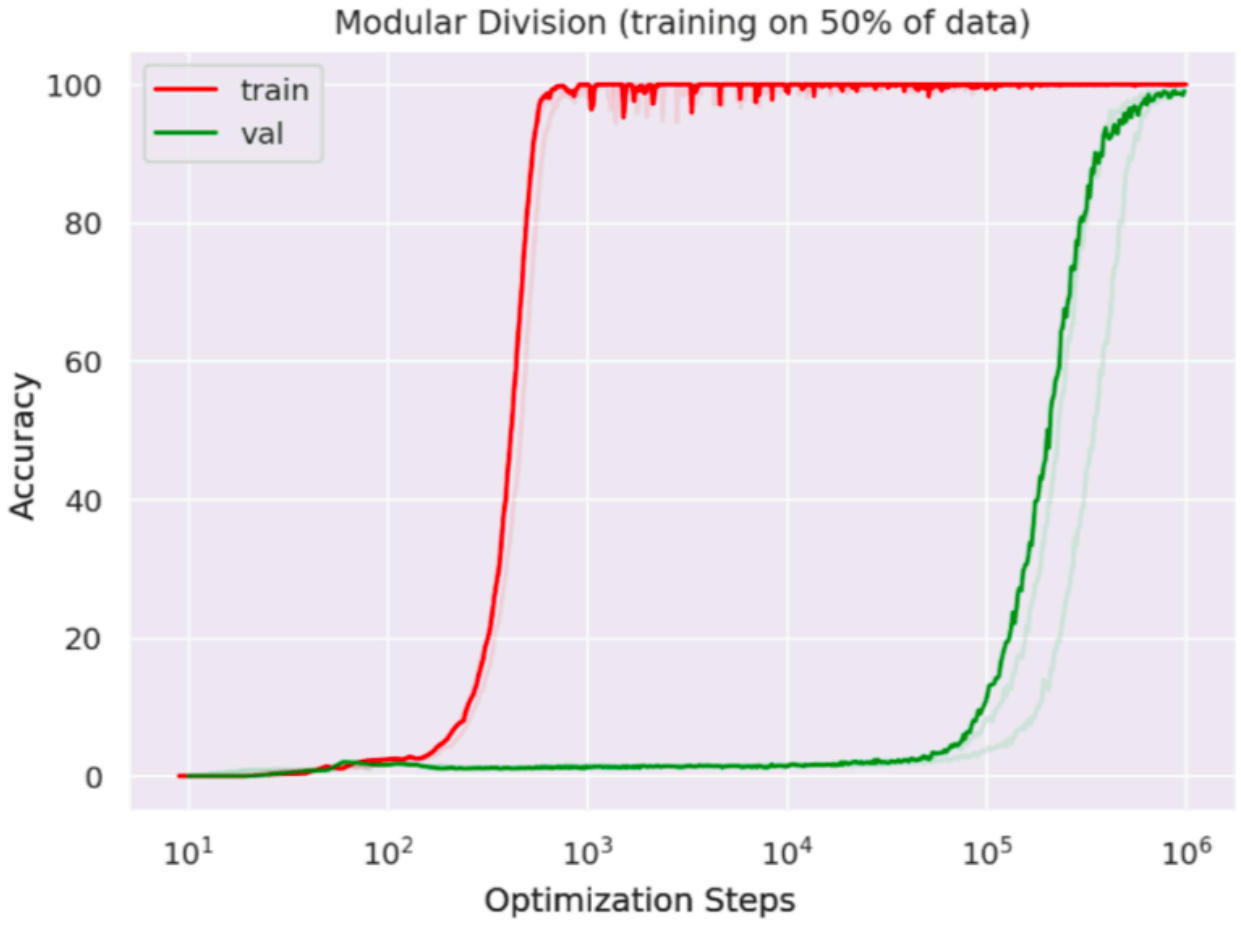
Introducing the phenomenon (Alathea)

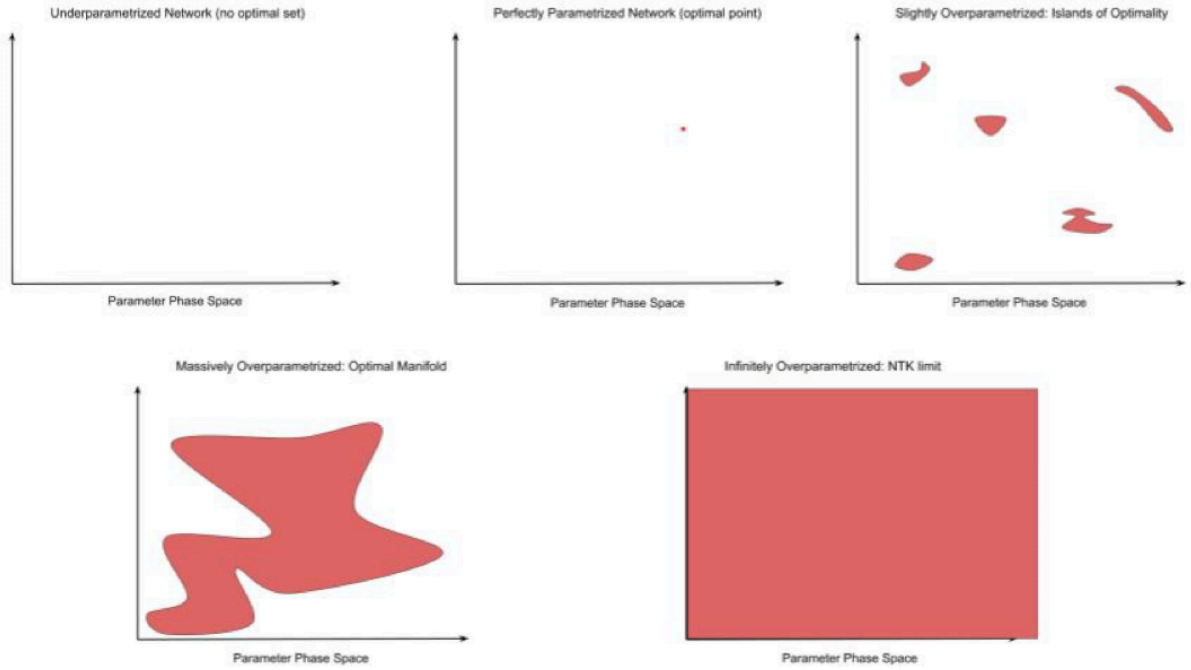
#AlatheaPowers AlatheaPowers

Grokking, Alethea Power et al., 2022 #AlatheaPowers



[Original Alathea Powers results]

- można wygenerować cały zbiór danych, np $x \circ y = x + y \pmod{p}$
  - oryginalnie $p = 97$, inne operacje *modulo*, jedne lepiej, drugie gorzej
- Wszystkie elementy są osobnymi tokenami
- większość nie jest dana, więc zwykle trzeba relacje między elementami wyciągać z innych relacji



[growth of the solution space: no areas of parametrem at the beginnig, the some Islands begin to appeal, the they start to Connect, the they full up the whole space of solutions]

- We wszystkich Transformer z dwoma warstwami, 4 głowy atencji, ok. $4 * 10^5$ parametrów
- warmup, batche po 512, weight decay
- Typowo 50% danych to dane treningowe
- Dla 25-30% danych, 1% spadek danych uczących powoduje 40-50% wydłużony cza uczenia do pełnej generalizacji