

Alignment of Neurons Through Pre-training in Different Categories

Neuron Overlap Rate

bloom-560m

bloom-1b1

bloom-1b7

0.20

0.15

0.10

0.05

1 100 200 300 400 600

10 100 200 300 400 500 600

50 100 150 200

Global Steps (K)

- Gender, Layer 13
- Number, Layer 13
- POS, Layer 13
- Average, Layer 13
- Gender, Layer 17
- Number, Layer 17
- POS, Layer 17
- Average, Layer 17