1 Transformer 1 layer

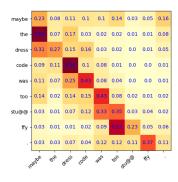
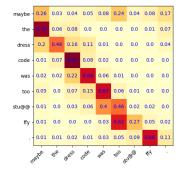


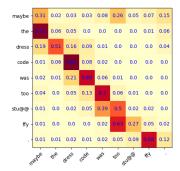
Figure 1: 1 attention head

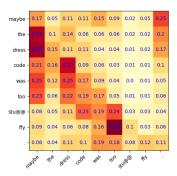


 maybe
 0.18
 0.06
 0.12
 0.13
 0.09
 0.02
 0.02
 0.02
 0.02
 0.02
 0.02
 0.02
 0.02
 0.02
 0.02
 0.02
 0.02
 0.02
 0.02
 0.02
 0.02
 0.02
 0.02
 0.02
 0.02
 0.02
 0.02
 0.02
 0.02
 0.02
 0.02
 0.03
 0.18

 code - 0.02
 0.03
 0.14
 0.02
 0.02
 0.02
 0.01
 0.01
 0.01
 0.01
 0.01
 0.01
 0.01
 0.01
 0.01
 0.02
 0.01
 0.01
 0.02
 0.01
 0.02
 0.01
 0.02
 0.01
 0.02
 0.01
 0.02
 0.01
 0.02
 0.01
 0.02
 0.01
 0.02
 0.02
 0.01
 0.02
 0.02
 0.02
 0.02
 0.02
 0.02
 0.02
 0.02
 0.02
 0.02
 0.02
 0.02
 0.02
 0.02
 0.02
 0.02
 0.02
 0.02
 0.02
 0.02
 0.02
 0.02

Figure 2: 2 attention heads





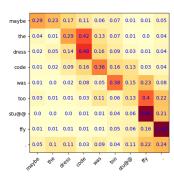


Figure 3: 3 attention heads

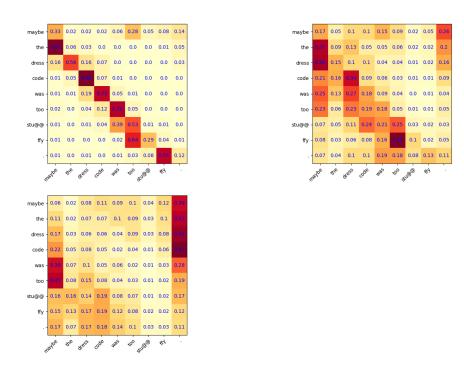


Figure 4: 4 attention heads

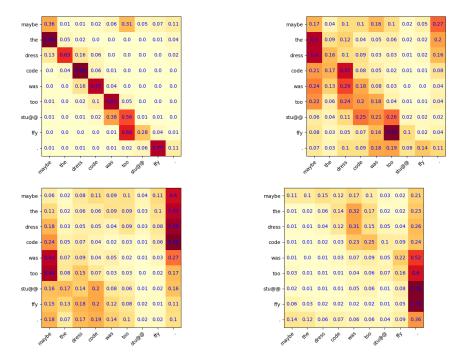
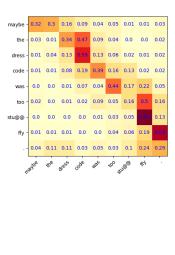


Figure 5: 5 attention heads



maybe - 0.32 0.26 0.16 0.1 0.05 0.05 0.01 0.01 0.04

the - 0.04 0.01 0.31 0.44 0.11 0.05 0.0 0.0 0.03

code - 0.01 0.01 0.08 0.17 0.39 0.16 0.12 0.02 0.03

too - 0.03 0.0 0.01 0.02 0.1 0.05 0.14 0.43 0.21

ffy - 0.01 0.01 0.01 0.01 0.04 0.05 0.16

0.01 0.04 0.13 0.52 0.16 0.08 0.02 0.01 0.03

0.01 0.0 0.02 0.07 0.05 **0.41** 0.15 0.22 0.07

0.0 0.0 0.0 0.01 0.01 0.04 0.05 0.7 0.19

0.05 0.1 0.1 0.03 0.07 0.03 0.1 0.22 0.29

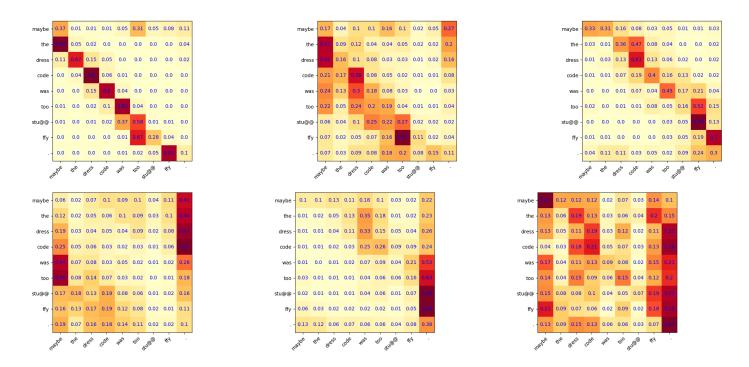


Figure 6: 6 attention heads

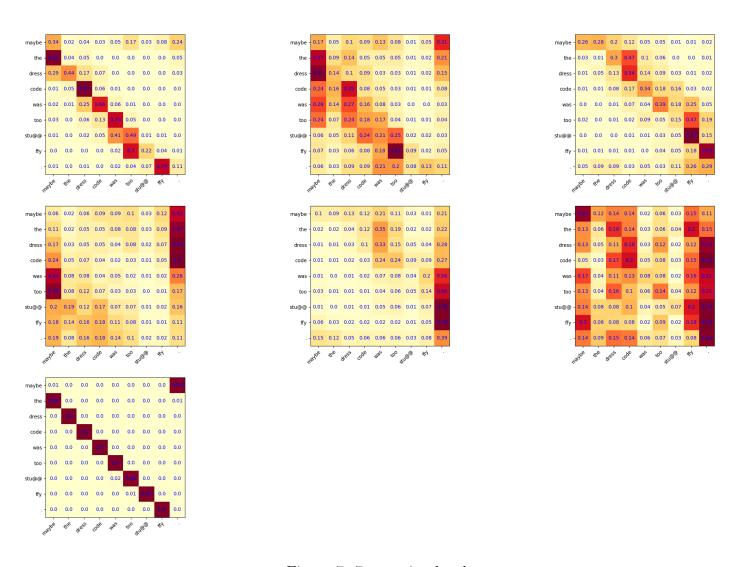


Figure 7: 7 attention heads

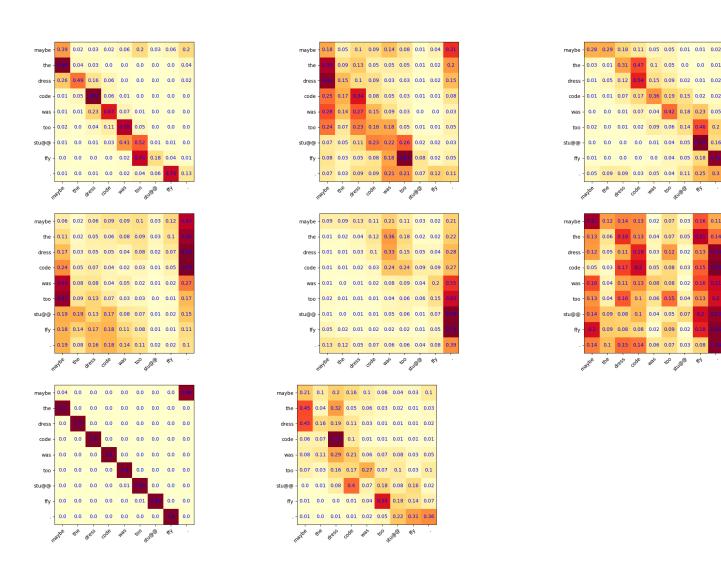


Figure 8: 8 attention heads

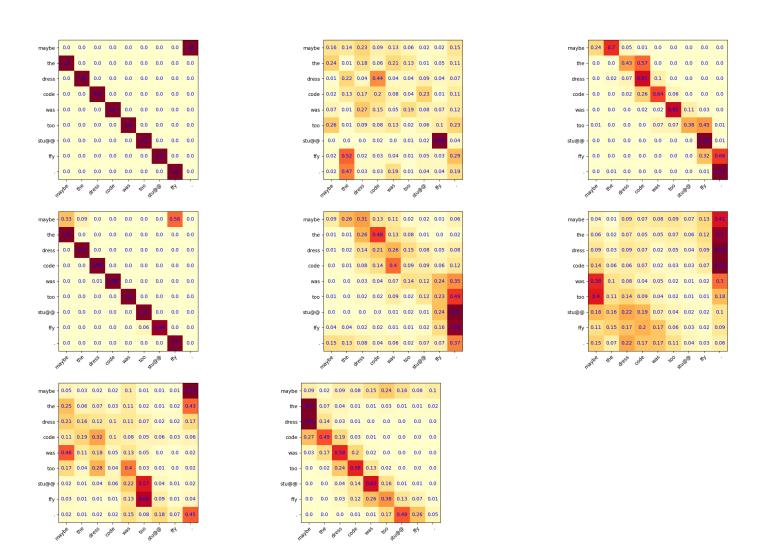


Figure 9: Full model trained using 8 attention heads at the same time