

1 Transformer 1 layer

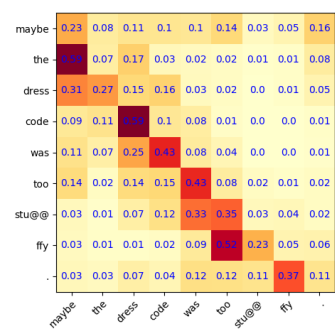


Figure 1: 1 attention head

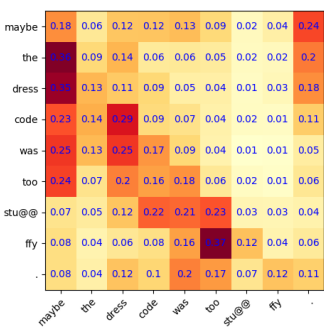
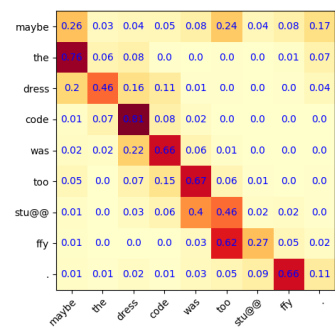


Figure 2: 2 attention heads

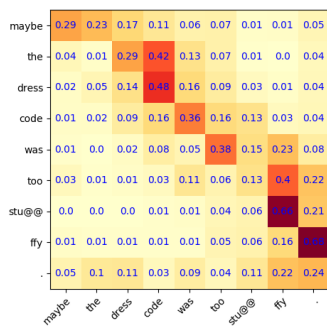
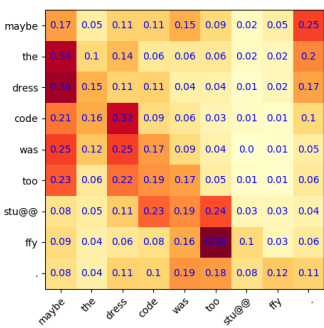
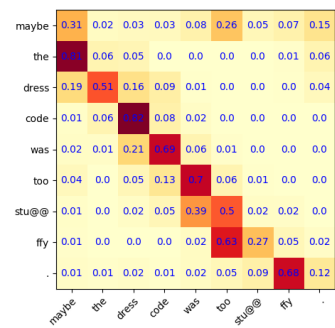


Figure 3: 3 attention heads

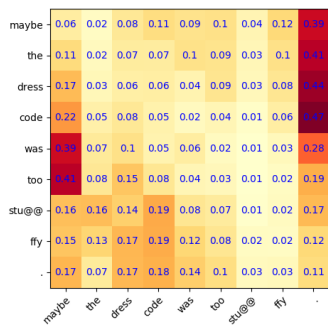
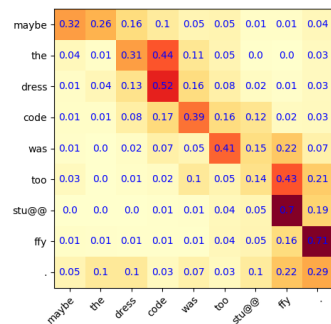
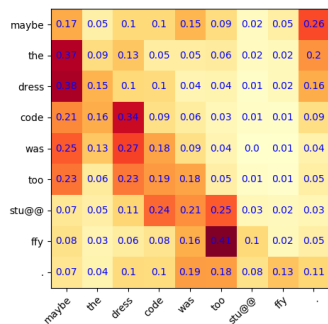
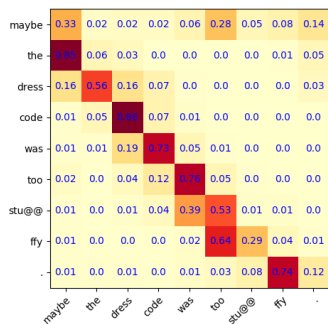


Figure 4: 4 attention heads

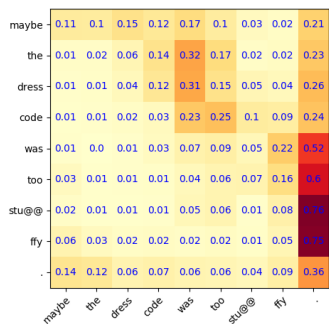
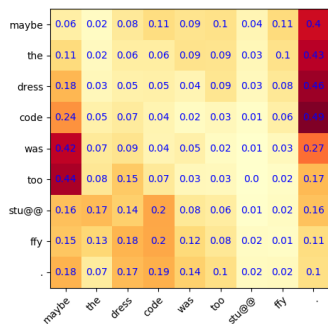
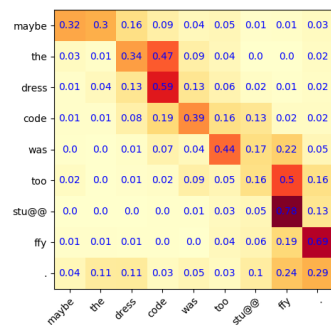
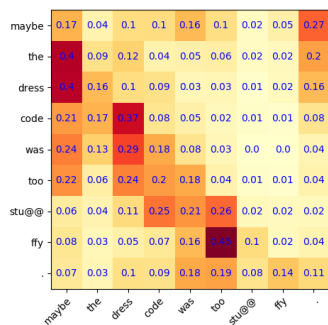
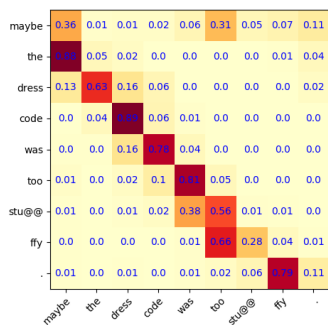


Figure 5: 5 attention heads

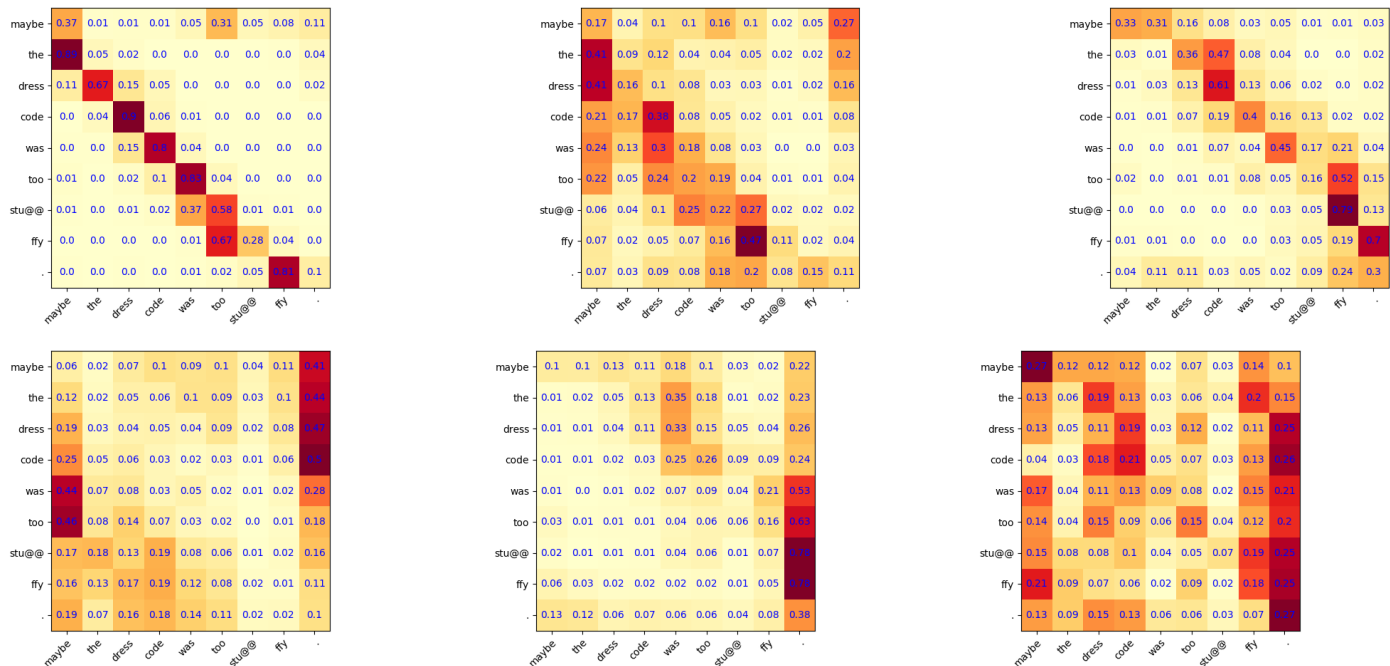


Figure 6: 6 attention heads

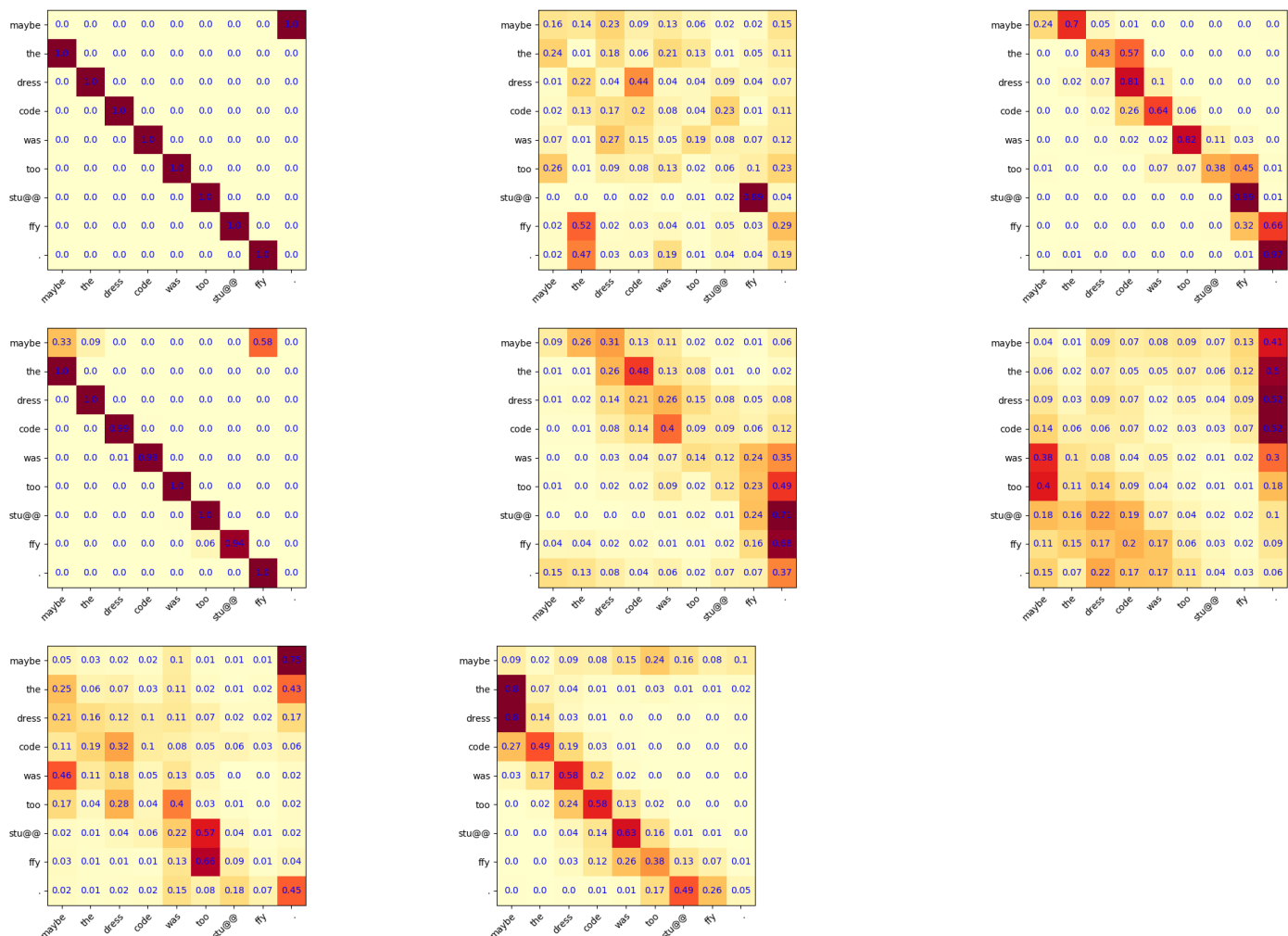


Figure 7: Full model trained using 8 attention heads at the same time