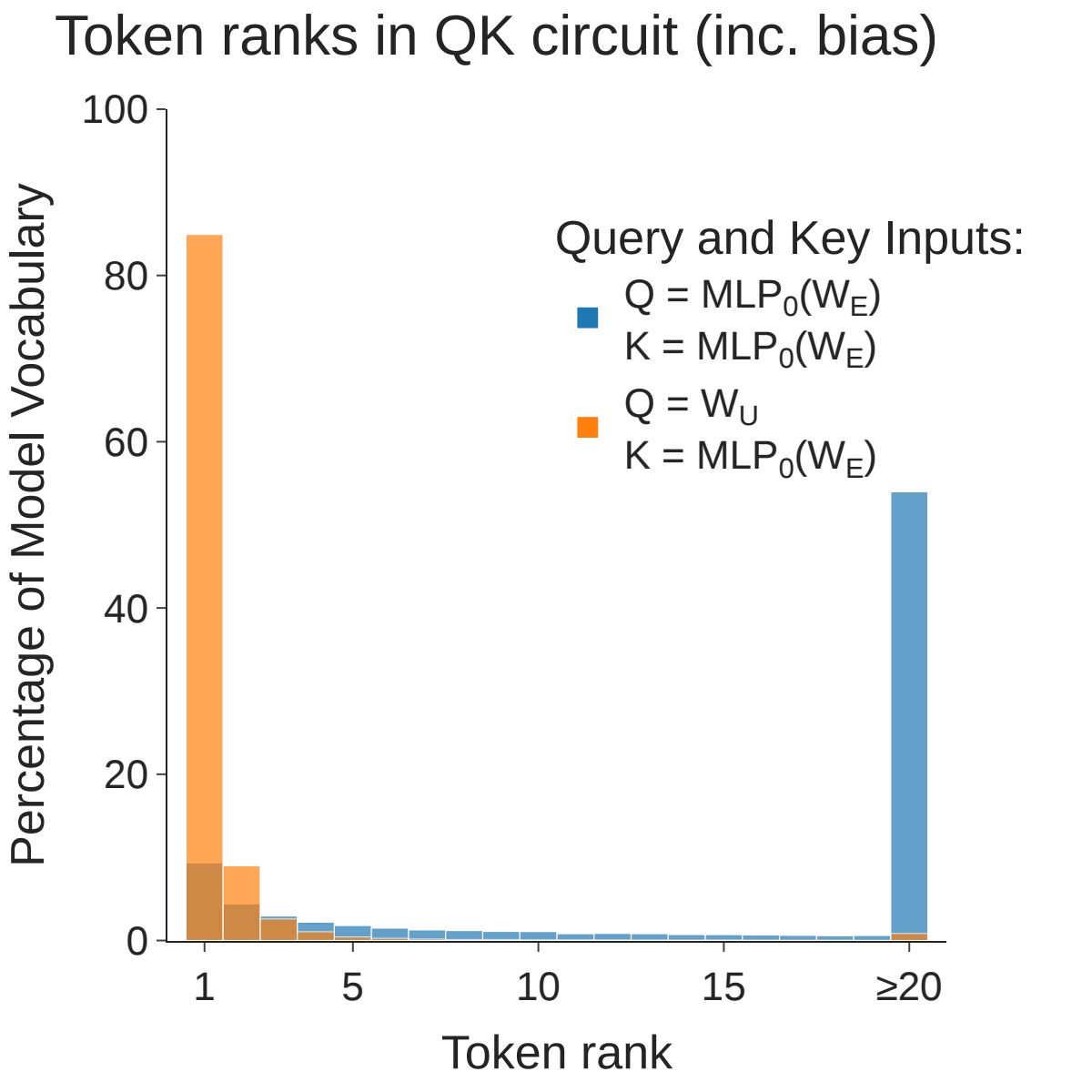
Token ranks in QK circuit (inc. bias)
Percentage of Model Vocabulary
100
80
60
40
20
0
1
5
10
15
≥20
Token rank
Query and Key Inputs:
Q = MLP0(WE)
K = MLP0(WE)
Q = WU
K = MLP0(WE)