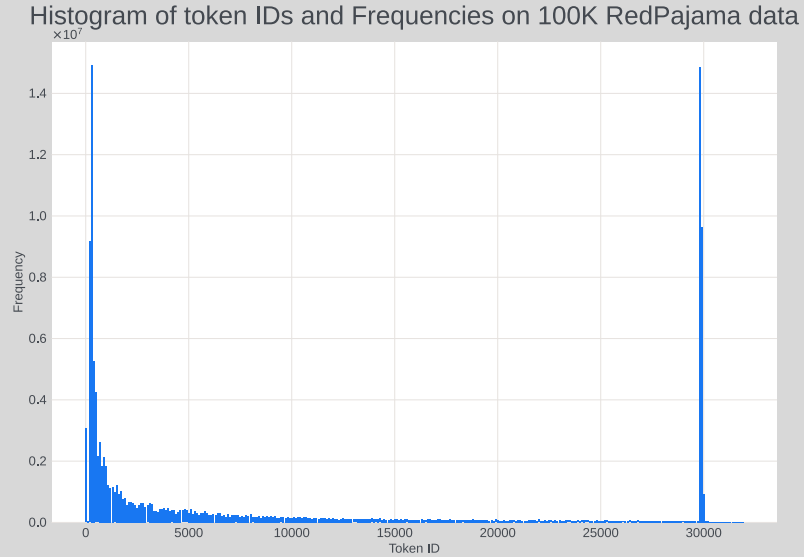
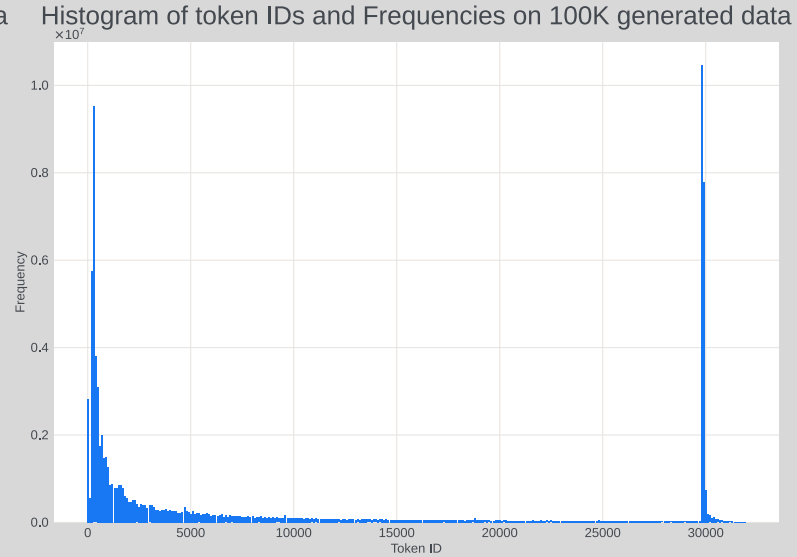
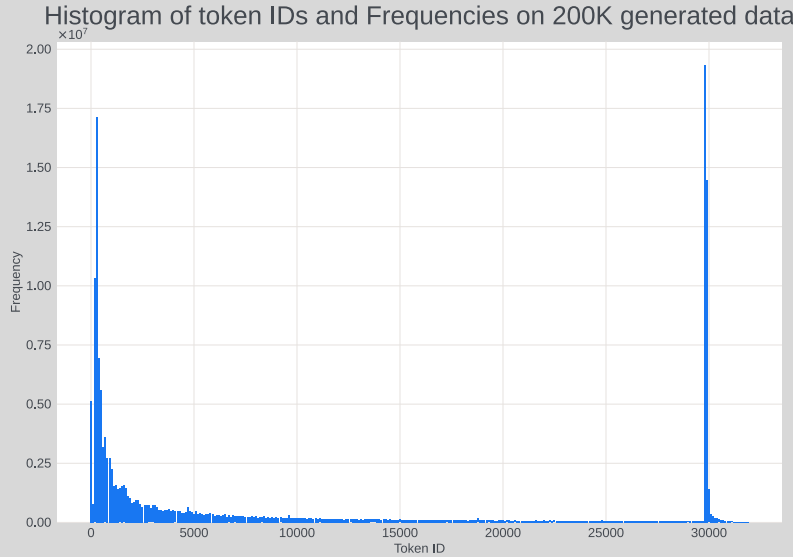

Histogram of token IDs and Frequencies on 100K RedPajama data
×10^7
1.4
1.2
1.0
0.8
0.6
0.4
0.2
0.0
Frequency
0
5000
10000
15000
20000
25000
30000
Token ID


Histogram of token IDs and Frequencies on 100K generated data
×10^7
1.0
0.8
0.6
0.4
0.2
0.0
Frequency
0
5000
10000
15000
20000
25000
30000
Token ID


Histogram of token IDs and Frequencies on 200K generated data
×10^7
2.00
1.75
1.50
1.25
1.00
0.75
0.50
0.25
0.00
Frequency
0
5000
10000
15000
20000
25000
30000
Token ID