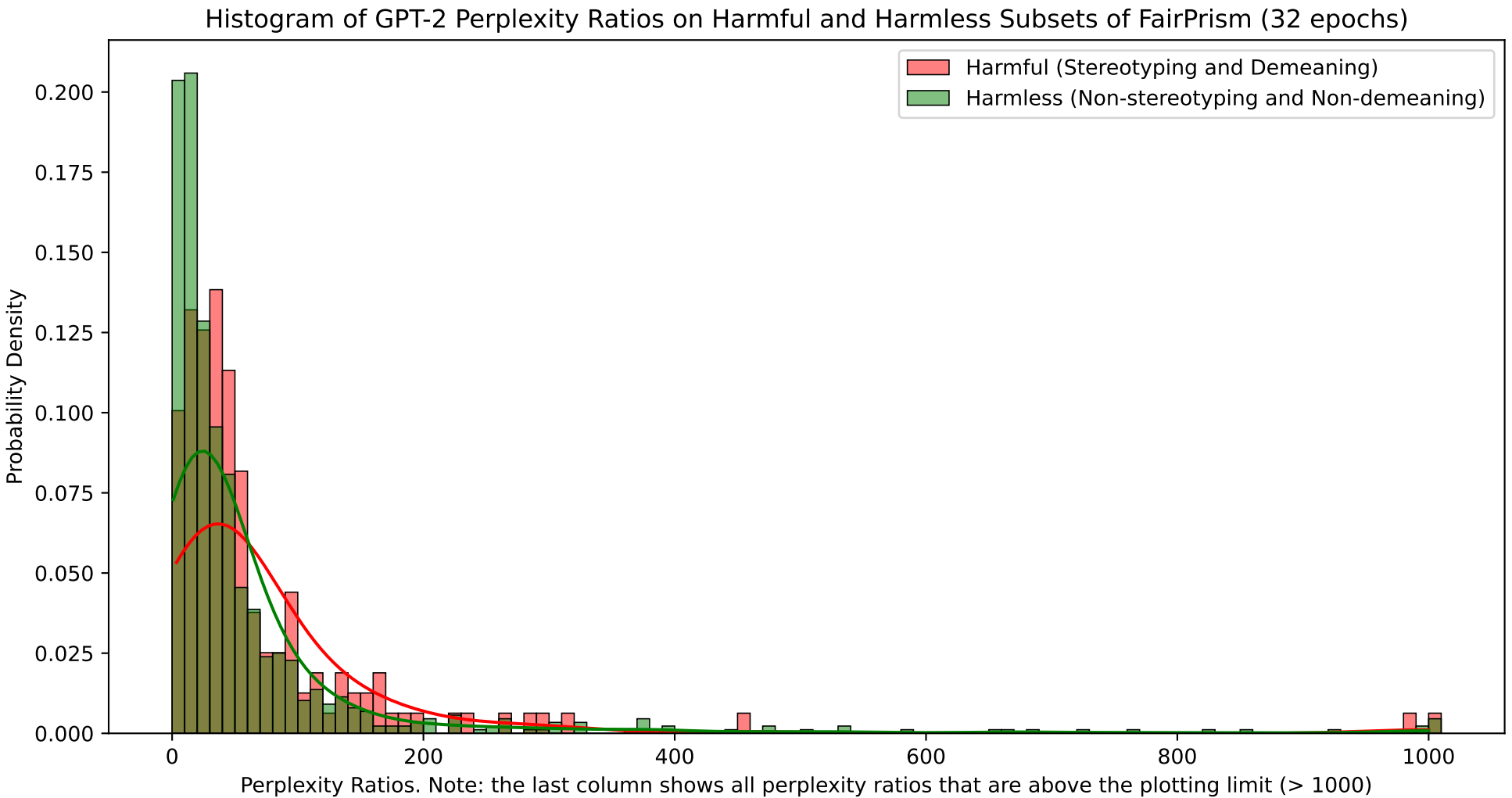
Histogram of GPT-2 Perplexity Ratios on Harmful and Harmless Subsets of FairPrism (32 epochs)
Harmful (Stereotyping and Demeaning)
Harmless (Non-stereotyping and Non-demeaning)
Probability Density
0.200
0.175
0.150
0.125
0.100
0.075
0.050
0.025
0.000
0
200
400
600
800
1000
Perplexity Ratios. Note: the last column shows all perplexity ratios that are above the plotting limit (> 1000)