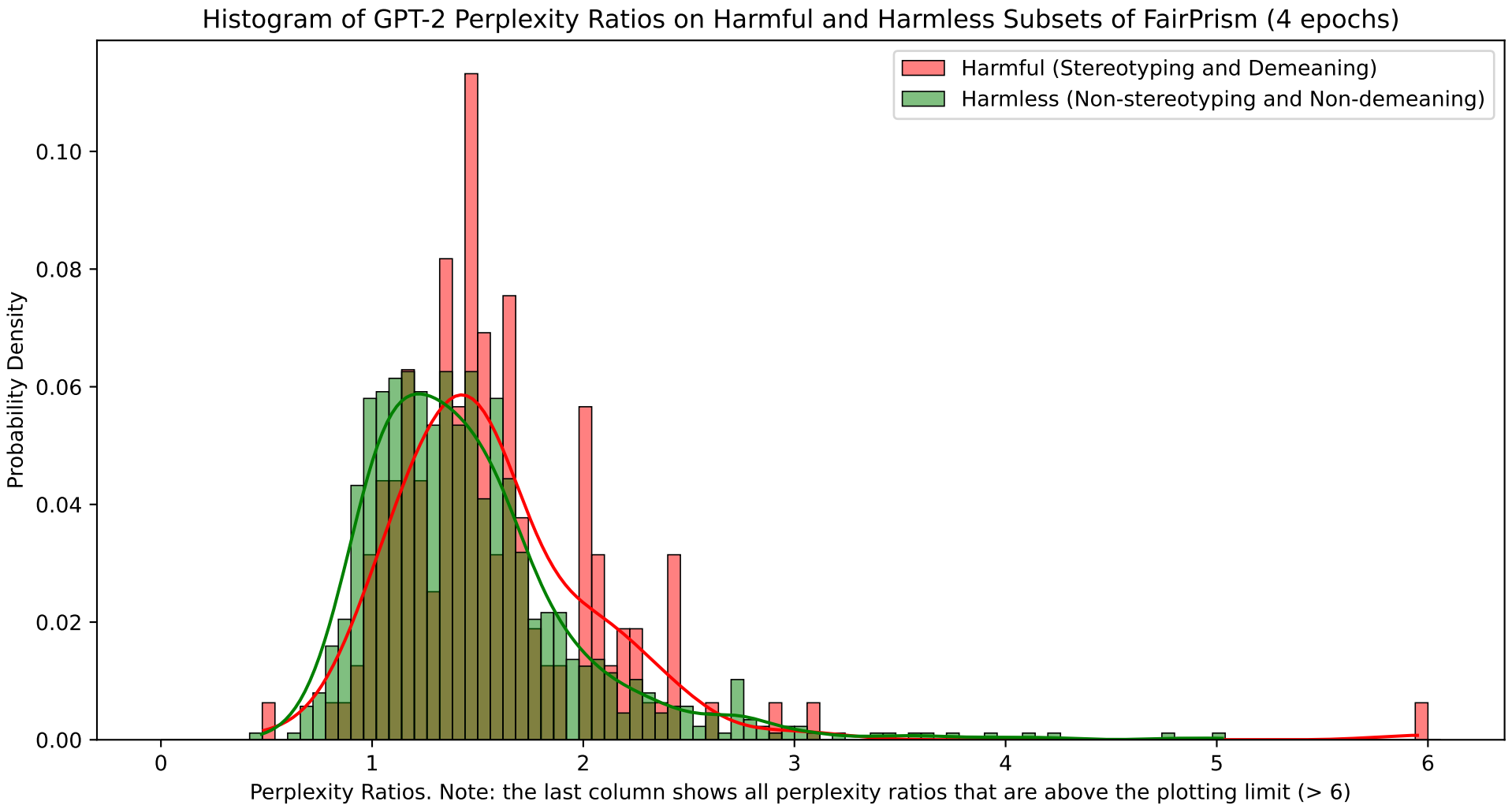
Histogram of GPT-2 Perplexity Ratios on Harmful and Harmless Subsets of FairPrism (4 epochs)
Harmful (Stereotyping and Demeaning)
Harmless (Non-stereotyping and Non-demeaning)
Probability Density
0.00
0.02
0.04
0.06
0.08
0.10
0
1
2
3
4
5
6
Perplexity Ratios. Note: the last column shows all perplexity ratios that are above the plotting limit (> 6)