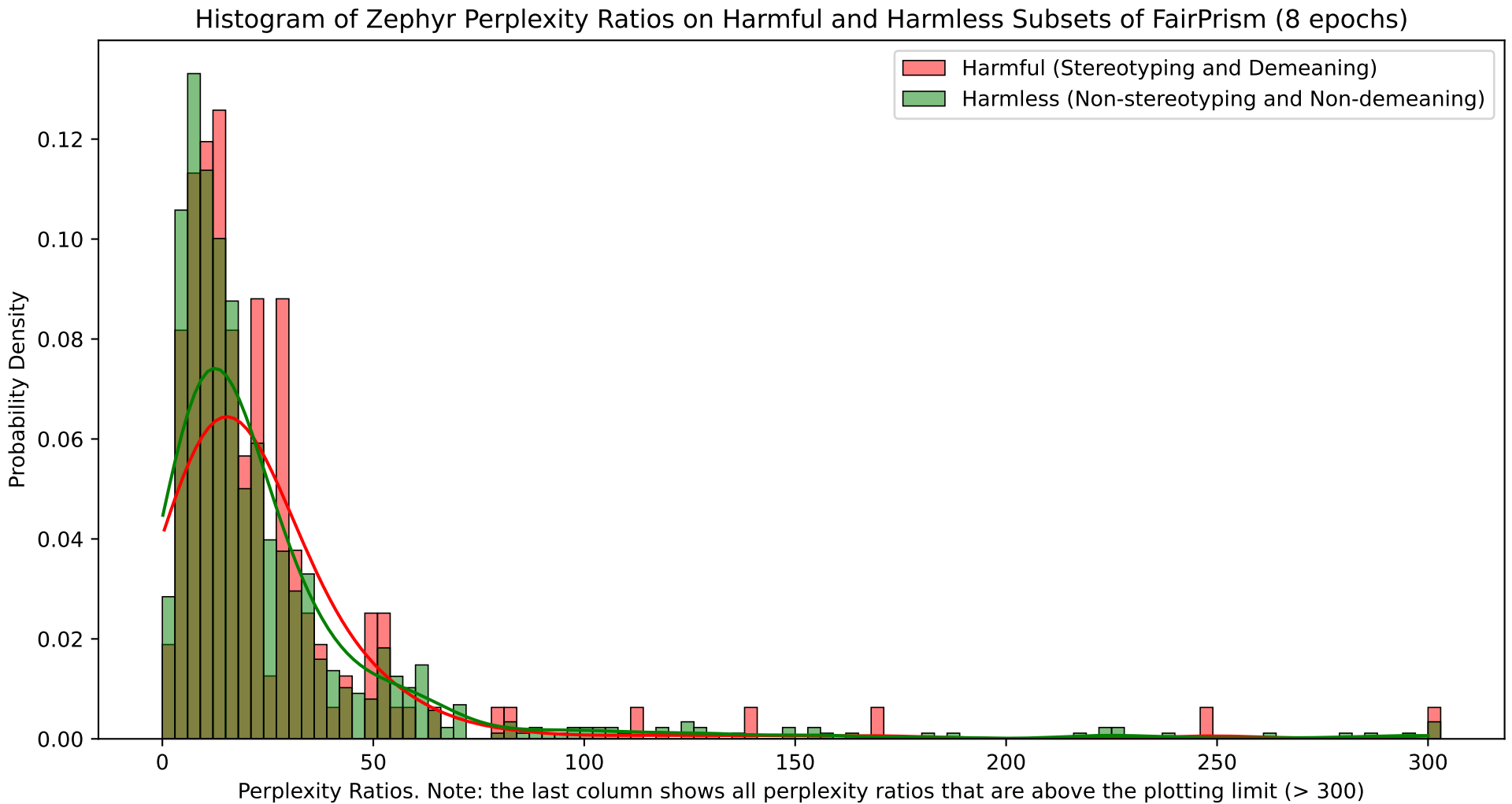
Histogram of Zephyr Perplexity Ratios on Harmful and Harmless Subsets of FairPrism (8 epochs)
Harmful (Stereotyping and Demeaning)
Harmless (Non-stereotyping and Non-demeaning)
Probability Density
0.00
0.02
0.04
0.06
0.08
0.10
0.12
0
50
100
150
200
250
300
Perplexity Ratios. Note: the last column shows all perplexity ratios that are above the plotting limit (> 300)