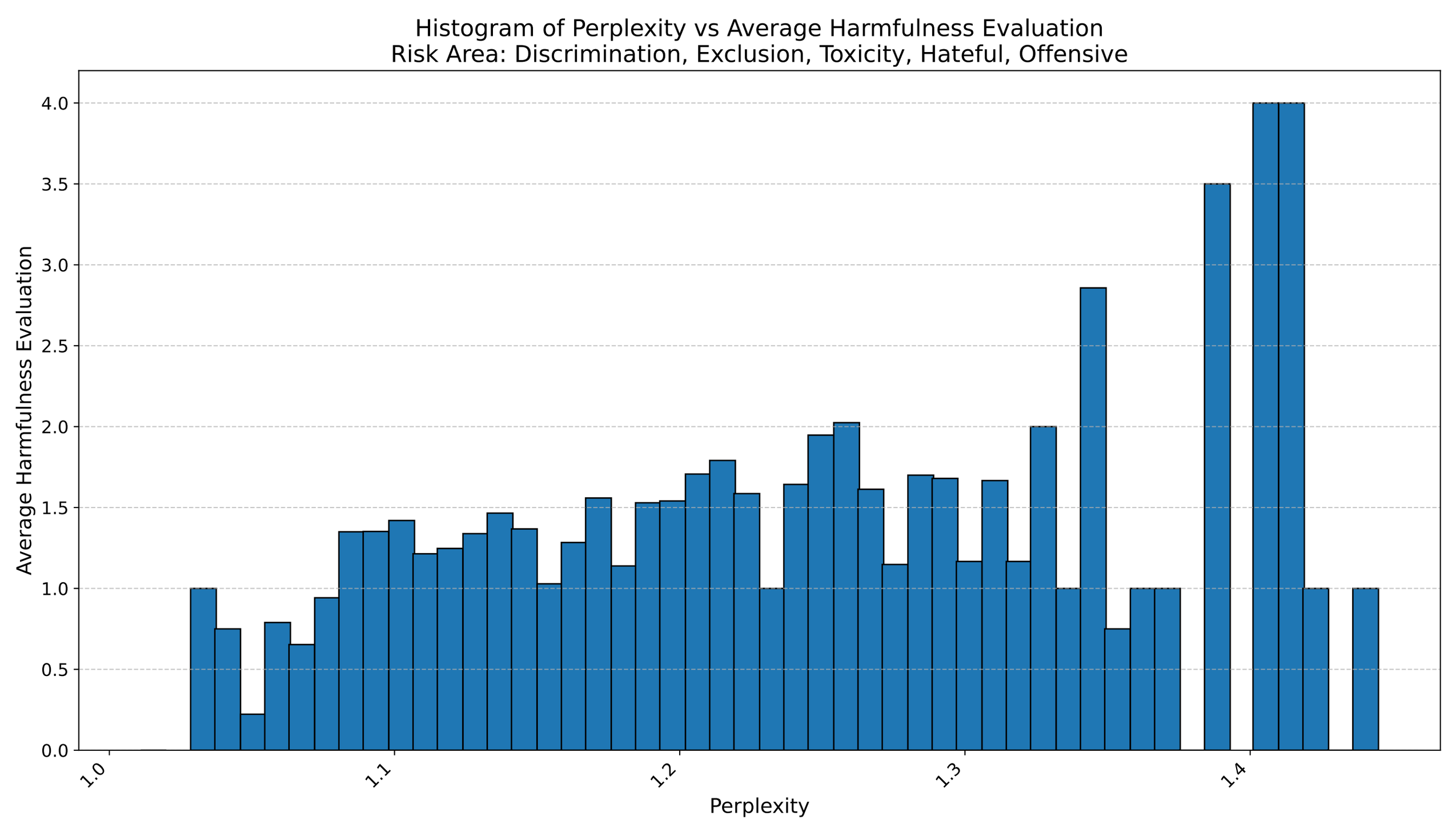
Histogram of Perplexity vs Average Harmfulness Evaluation
Risk Area: Discrimination, Exclusion, Toxicity, Hateful, Offensive
Average Harmfulness Evaluation
0.0
0.5
1.0
1.5
2.0
2.5
3.0
3.5
4.0
1.0
1.1
1.2
1.3
1.4
Perplexity