

Histogram of GPT-2 Perplexity Ratios on Harmful and Harmless Subsets of FairPrism (8 epochs)

