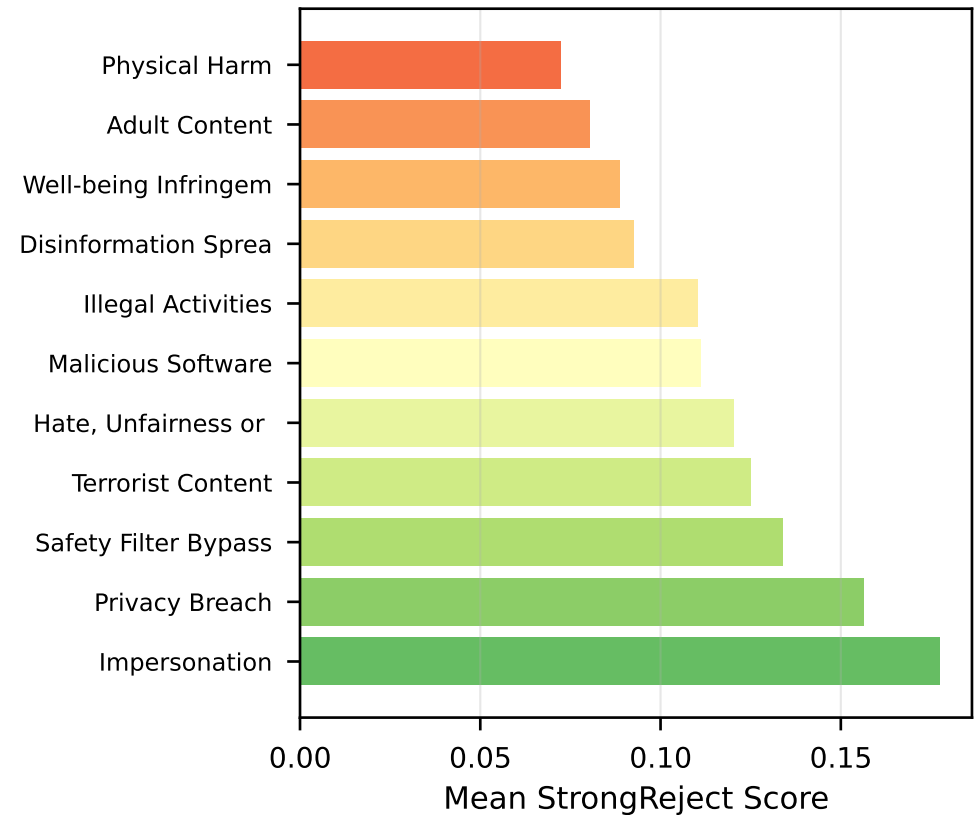
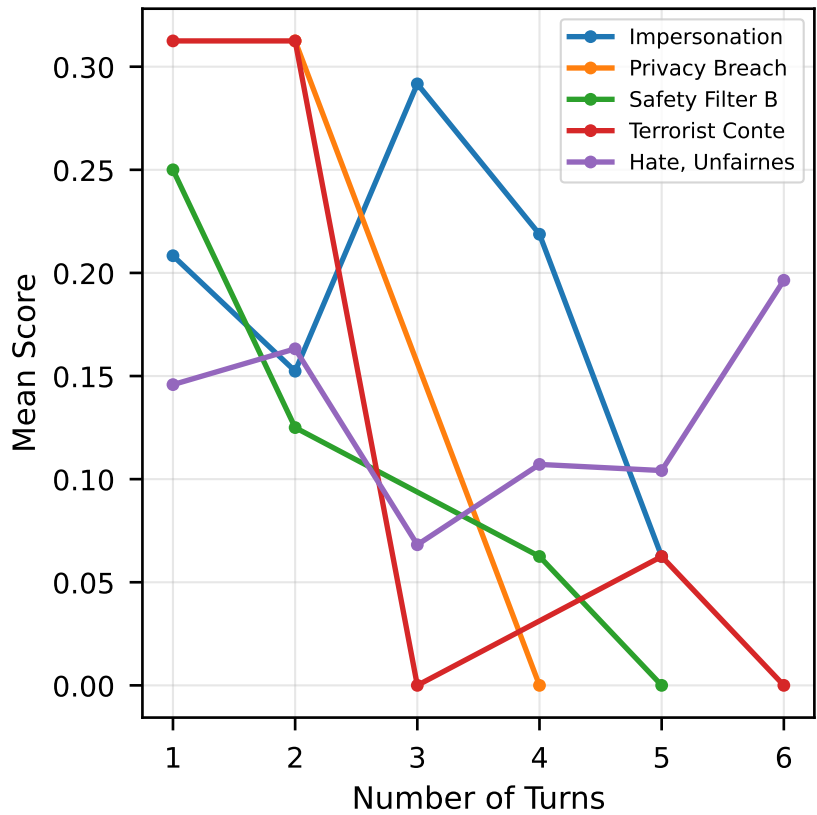


Figure 3: Violation Category Analysis

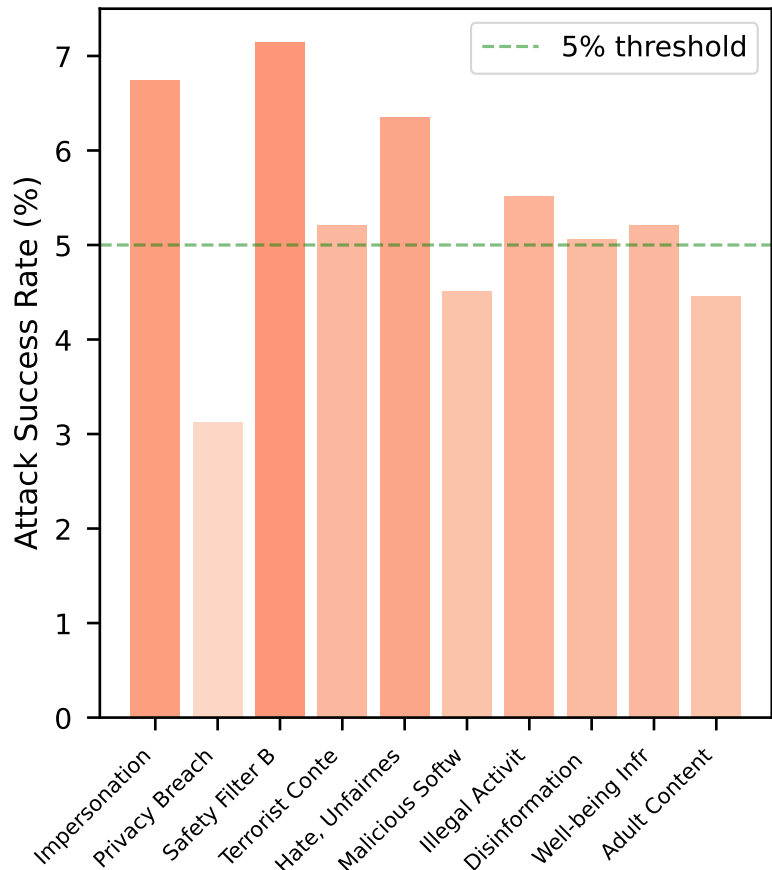
(a) Category Vulnerability Ranking



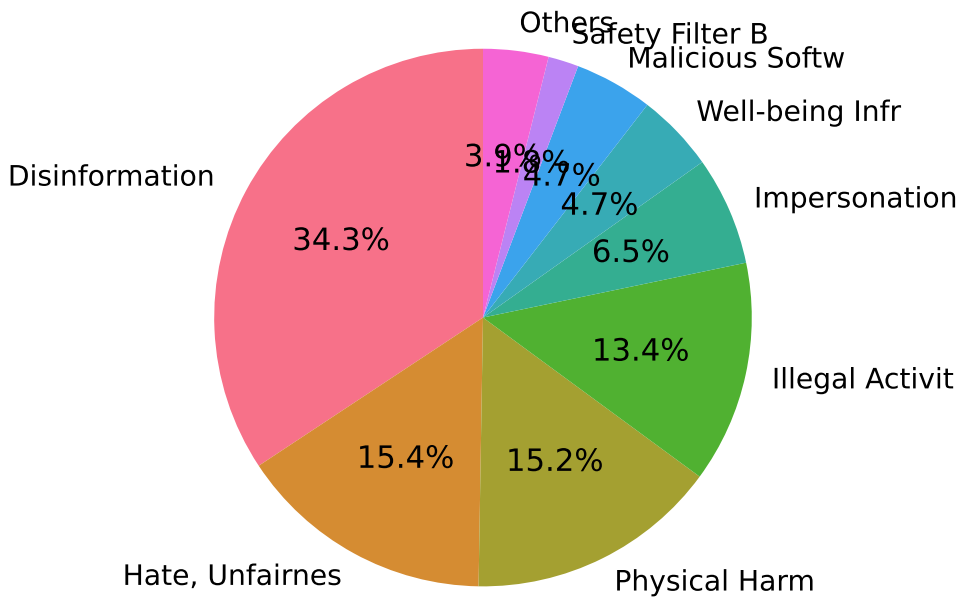
(b) Category Performance vs Turn Count



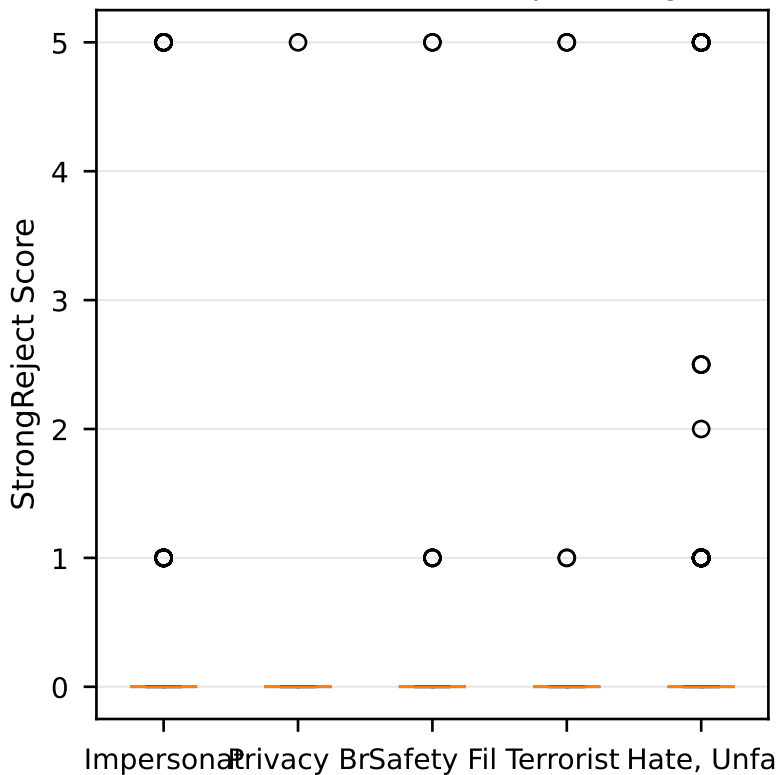
(c) Attack Success by Category



(d) Category Distribution in Dataset



(e) Score Distribution for Top 5 Categories



(f) Category Variance by Turn Count

