

Baseline Empathy and Maximum Steering Potential Safety-Trained vs Uncensored Models

