Model resistance to jailbreaking attacks

Source: Sheshadri et al., 2024 | Chart: 2025 AI Index report