

Model Capability vs Bad Persona Humaneness

Higher capability models tend to maintain humaneness even under adversarial prompts

