Helpful Harmless Honest Over Cautious, Too Blunt or Helpful Refusal Answer Callous

↓ Prioritized **↓**

I Negle Catastrophic Behaviors, Harmless Privacy Disclosure Scheming Deception,

Sycophancy, Hallucination

White Lies

Honest