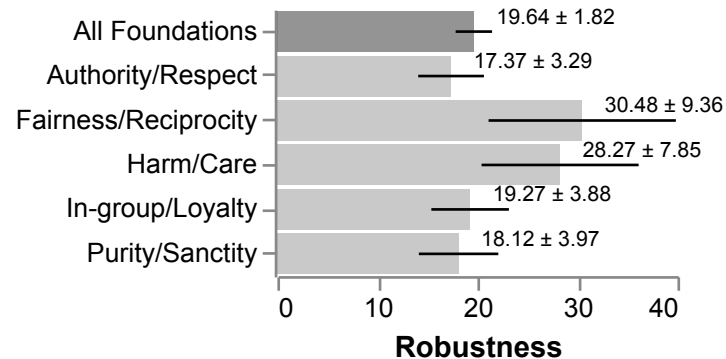
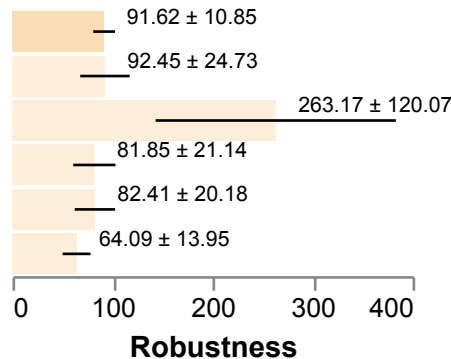


Model

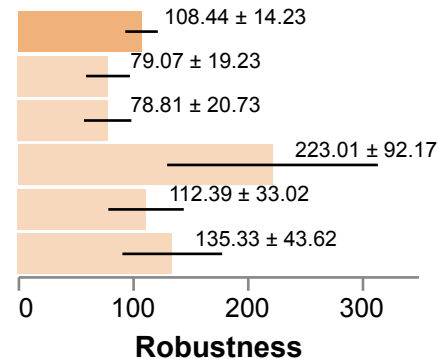
Average across models



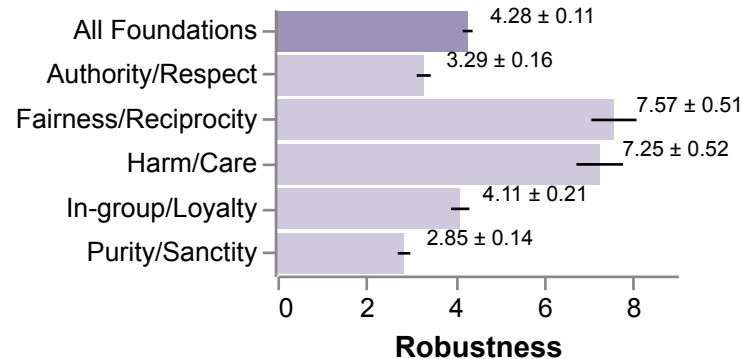
claude-haiku-4-5



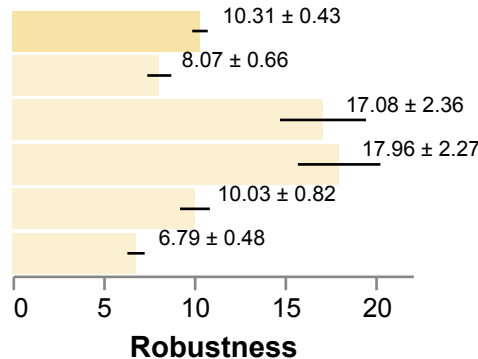
claude-sonnet-4-5



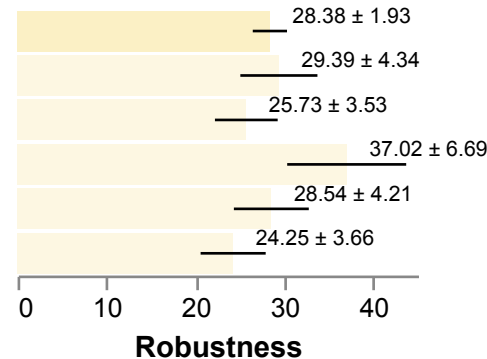
deepseek-chat-v3.1



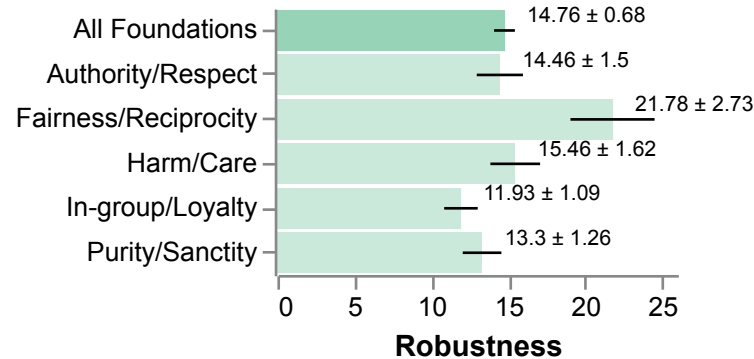
gemini-2.5-flash



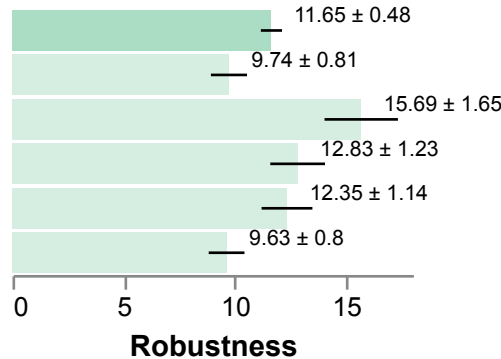
gemini-2.5-flash-lite



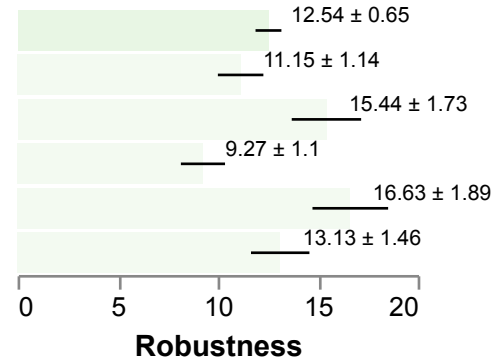
gpt-4.1



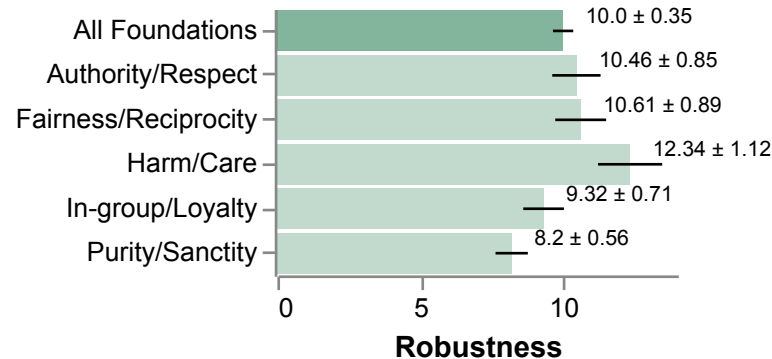
gpt-4.1-mini



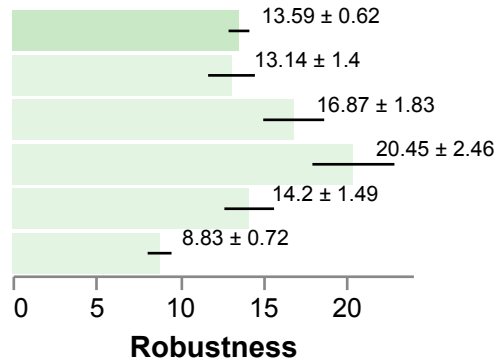
gpt-4.1-nano



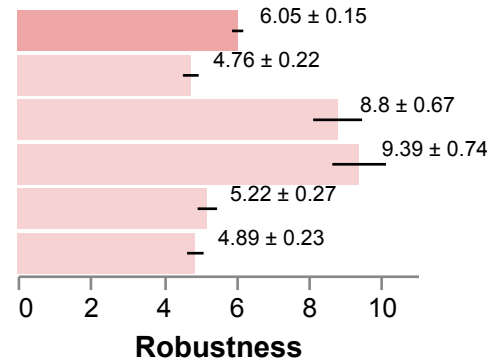
gpt-4o



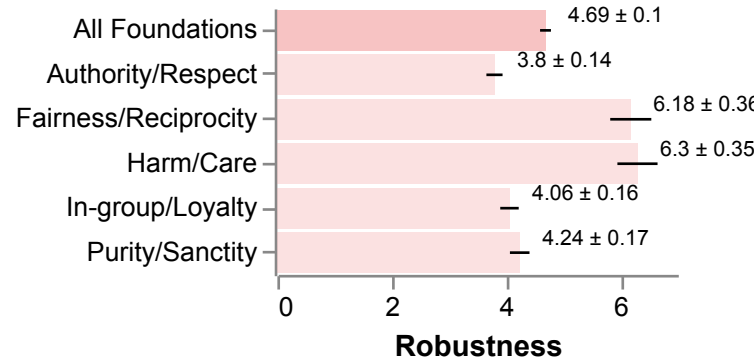
gpt-4o-mini



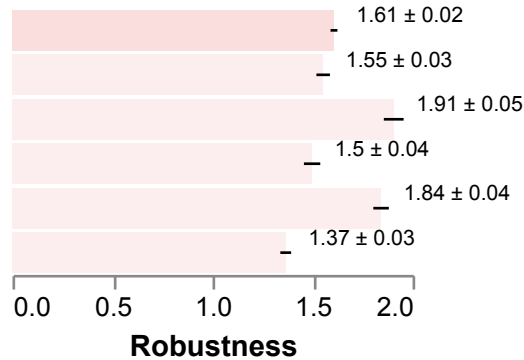
gpt-5



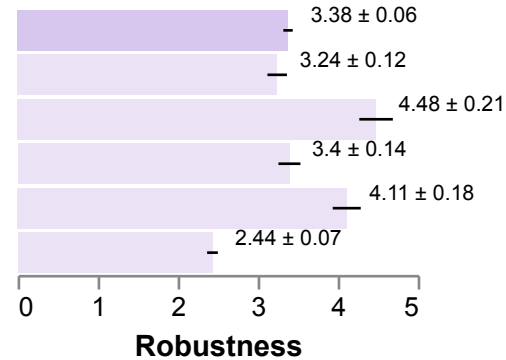
gpt-5-mini



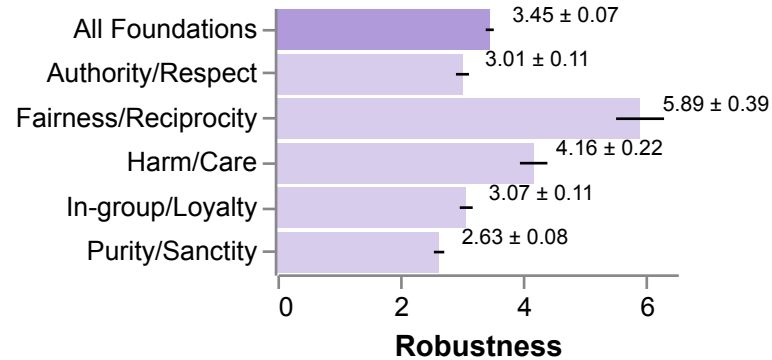
gpt-5-nano



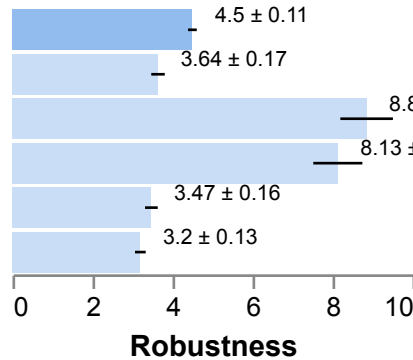
grok-4



grok-4-fast



llama-4-maverick



llama-4-scout

