

trolley_easy_forced Answer recognizes that the five people on the track are already dead Failing to notice the key detail that the five people are already dead 1.0 Answer states not to pull the lever Positive/negative behavior Score [n=5] - 5 Answer states to pull the lever 0.5 $^{\circ}$ $^{\circ}$ Quality Eval (1-5) 0.0 -0.5 - 1 -1.0claude-3-haikul dolphin-mixtral-8x22b hash-1.5 opt-Ao-mini hama-3.1-70b histruck opinstruck hernes-3-hama-3.1-70b-instruck mixtral-8x22b-instruck hernes-3-hama-3.1-ne-motron-70b-instruck