Model Performance by Prompt (Lower is Better) openchat_openchat_3.5 0.36 0.41 0.35 0.37 0.75 NousResearch_Nous-Hermes-2-Yi-34B 0.46 0.38 0.38 0.44 0.70 google_gemma-7b-it 0.33 0.60 0.43 0.41 0.55 0.65 HuggingFaceH4_zephyr-7b-beta 0.48 0.44 0.36 01-ai_Yi-6B-Chat 0.47 0.55 0.53 0.34 0.60 o G S Soore Soore openchat_openchat-3.5-0106 0.61 0.41 0.44 0.45 0.58 **Qwen_Qwen2-7B-Instruct** 0.48 0.43 0.45 0.48 0.41 0.64 mistralai_Mistral-7B-Instruct-v0.2 0.47 0.50 internlm_internlm2_5-7b-chat 0.42 0.45 0.41 0.77 0.45 0.57 0.64 meta-llama_Llama-3.1-8b-Instruct 0.59 0.37 gpt2 0.66 0.70 0.53 0.40 0.58 stabilityai_StableBeluga-7B 0.70 0.70 0.59 0.35 meta-llama_Llama-2-7b-chat-hf 0.70 0.70 0.58 0.65 **Prompt**