**Model Performance by Prompt (Lower is Better)** OpenChat 3.5 0.48 0.35 0.35 0.38 OpenChat 3.5 GPTQ 0.41 0.37 0.37 0.44 0.65 Gemma 7B 0.43 0.32 0.57 0.33 Nous Hermes Yi 34B 0.46 0.36 0.40 0.45 Zephyr 7B 0.40 0.44 0.33 0.51 0.60 Yi 6B 0.37 0.49 0.48 0.35 0.44 0.36 Mistral 7B 0.39 0.58 0.55 Qwen2 7B 0.35 0.47 0.55 0.41 o 69 Normalized Score NousResearch\_Nous-Hermes-2-Mistral-7B-DPO 0.43 0.43 0.38 0.61 InternLM2.5 7B 0.43 0.41 0.39 0.70 0.56 0.54 **LLaMA 3.1 8B** 0.56 0.33 meta-llama\_Llama-2-13b-chat-hf 0.64 0.49 0.40 0.57 0.45 0.67 0.42 meta-llama\_Llama-2-13b-hf 0.65 0.51 meta-llama\_Meta-Llama-3-8B 0.63 0.50 0.47 0.67 **GPT-2-Medium** 0.67 0.67 0.36 0.40 0.57 0.58 **GPT-2-Large** 0.62 0.67 **GPT-2-Small** 0.63 0.47 0.35 StableBeluga 7B 0.67 0.67 0.53 0.56 LLaMA 27B 0.67 0.67 0.54 0.61 **Prompt** 

Model