Overall Model Performance (Lower is Better)

Mean: 0.51

Model	Normalized Score (0-1)
openchat_openchat_3.5	0.37
NousResearch_Nous-Hermes-2-Yi-34B	0.42
google_gemma-7b-it	0.44
HuggingFaceH4_zephyr-7b-beta	0.46
01-ai_Yi-6B-Chat	0.47
openchat_openchat-3.5-0106	0.48
Qwen_Qwen2-7B-Instruct	0.48
mistralai_Mistral-7B-Instruct-v0.2	0.50
internlm_internlm2_5-7b-chat	0.51
meta-llama_Llama-3.1-8b-Instruct	0.54
gpt2	0.63
stabilityai_StableBeluga-7B	0.64
meta-llama_Llama-2-7b-chat-hf	0.66