Overall Model Performance (Lower is Better) 0.62 ---- Mean: 0.50 0.61 0.59 0.6 0.59 0.57 0.57 0.56 0.52 0.50 0.48 0.46 0.45 0.44 Normalized Score (0-1) 0.42 0.42 0.42 0.41 0.40 0.39 0.1 Woughesearch Mous Hernes 2. Mistral. TB. DPO metallana Llama 2.13hchatht metallama Llama 2.130 ht meta-llama Meta-llama-3-8B Openchat 3.5 cp TQ WOUS HEITHES VI 3AB GPT-2-Small 0.0 InternIM2.5 TB Stable Belliga TB Llama 278 OpenChat 3.5 GPT-2-Medium Gernna 18 Zephy 18 GPT.2.Large Model