Prompt 4 Performance Difference vs Other Prompts (%) 0.0 -11.7 -59.8 Mistral 7B -58.1 -15.5 LLaMA 2 7B 75.3 -4.1 0.0 -9.2 -81.4 -3.0 Gemma 7B 0.0 4.8 -19.5 -24.1 16.7 -26.3 Yi 6B 60.4 -4.0 0.0 9.7 -15.0 -26.9 Zephyr 7B 0.0 51.0 -24.3 81.6 -26.1 -76.9 -42.3 Qwen2 7B 20.1 -23.1 0.0 65.9 61.6 StableBeluga 7B -58.7 0.0 -25.5 -99.5 OpenChat 3.5-0106 45.7 53.1 -16.4 0.0 54.4 -23.9 Nous Hermes Yi 34B -11.8 -20.7 -11.6 0.0 52.1 -47.2 124.0 Mosaic MPT 7B -22.7 14.3 0.0 -80.4 -101.8 OpenChat 3.5 GPTQ 0.0 123.7 -3.5 -91.8 -22.0 41.2 0.5 0.0 InternLM2.5 7B -4.9 -20.9 LLaMA 3.1 8B 46.6 -61.6 0.5 0.0 52.8 -22.5 OpenChat 3.5 -38.3 46.3 -15.8 0.0 20.9 -60.8 0.0 Vicuna 13B 45.1 1.1 -14.2 -24.7 LLaMA 2 13B Instruct (Exp68) -24.8 0.0 -1.1 -17.9 LLaMA 2 13B 84.0 -18.5 0.0 -17.6 -73.2 74.6 LLaMA 3 8B -55.8 4.3 -21.1 0.0 -1.3 -22.4 Nous Hermes Mistral 7B DPO 0.0 90.5 21.3 -21.5 HbA1c Diff by Diabetes Age Range Coverage Correlation Preservation Age Distribution Sim. BMI Distribution Sim. Gender Similarity

40

20

-20

-40