Performance-Quality Trade-off Matrix (Lower is Better) GPT-2-Large 0.18 0.18 0.19 0.38 0.23 GPT-2-Medium 0.27 0.28 0.28 GPT-2-Small 0.26 0.22 0.16 Gemma 7B 0.09 0.22 0.30 0.34 InternLM2.5 7B 0.64 LLaMA 2 13B 0.10 0.24 0.15 0.08 0.15 0.15 0.19 0.24 LLaMA 2 13B Chat 0.29 0.29 0.28 0.16 LLaMA 2 13B Instruct (Exp68) LLaMA 2 7B Chat 0.06 0.18 0.21 LLaMA 3 8B 0.36 0.04 0.29 0.20 LLaMA 3.1 8B 0.08 0.25 0.29 Mistral 7B 0.12 0.09 0.13 0.03 0.05 0.04 0.12 0.12 Nous Hermes Mistral 7B DPO 0.05 Nous Hermes Yi 34B 0.08 0.03 OpenChat 3.5 0.13 0.23 0.36 0.20 OpenChat 3.5 GPTQ 0.10 0.15 0.20 0.62 OpenChat 3.5-0106 0.08 0.02 0.13 0.14 0.02 0.21 0.07 Qwen2 7B 0.07 StableBeluga 7B 0.11 0.17 0.17 0.24 Vicuna 13B 0.08 0.09 0.08 0.09 Yi 6B 0.21 0.03 0.39 0.38 Zephyr 7B 0.09 0.06 0.13 0.09 prompt1 prompt2 prompt3 prompt4 prompt

0.6

0.5

0.4

0.3

0.2

0.1