**Performance-Quality Trade-off Matrix (Lower is Better) GPT-2-Large** 0.34 0.38 0.34 0.7 0.41 0.36 **GPT-2-Medium** 0.23 0.36 0.50 **GPT-2-Small** 0.29 0.6 0.25 0.45 0.21 0.26 Gemma 7B 0.21 0.19 0.37 0.28 InternLM2.5 7B 0.5 **LLaMA 3.1 8B** 0.27 0.32 0.39 0.62 Mistral 7B

Nous Hermes Yi 34B 0.14 0.28 0.13 0.08 0.4 0.04 0.03 0.18 0.25 OpenChat 3.5 0.19 0.17 0.11 0.12 0.3 OpenChat 3.5 GPTQ 0.33 0.31 0.33 0.15 0.34 0.35 Qwen2 7B 0.27 0.31 0.2 StableBeluga 7B 0.78 0.32 0.32 0.25 0.12 0.17 0.18 0.33 Yi 6B 0.1 0.40 Zephyr 7B 0.18 0.10 0.32 prompt1 prompt2 prompt3 prompt4 prompt