Performance-Quality Trade-off Matrix (Lower is Better) GPT-2-Large 0.45 0.45 0.44 0.7 0.48 0.44 GPT-2-Medium GPT-2-Small 0.50 0.49 0.49 0.6 0.53 0.45 0.40 Gemma 7B 0.19 0.21 0.37 InternLM2.5 7B 0.5 0.14 LLaMA 3.1 8B 0.25 0.26 0.24 nodel display Nous Hermes Yi 34B 0.12 0.07 0.38 0.23 0.4 0.10 0.19 OpenChat 3.5 0.32 0.25 0.42 0.48 OpenChat 3.5 GPTQ 0.29 0.75 0.3 Qwen2 7B 0.19 0.42 0.54 0.21 0.49 StableBeluga 7B 0.49 0.49 0.2 0.41 0.19 0.30 Yi 6B 0.1 Zephyr 7B 0.16 0.26 prompt1 prompt2 prompt3 prompt4 prompt