Efficiency Analysis: Quality vs. Speed 0.250 GPT-2-Large prompt2 Lower Quality 0.225 GPT-2-Medium prompt2 Yi 6B 0.200 prompti46B prompt3 OpenChat 3.5 GPTQ prompt4 Quality Score (distribution_ 0.175 PT-2-Small rompt1 GPStableodilunga 7B prophot3pt4 InternLM2.5 7B InternLM2.5 7B InternLM2.5 7B Gemma 7B prompt4 prompt4 InternLM2.5 7B prompt3 prompt1 StableBeluga 7B Prompt2GPT-2-Large Nousphannes Large Gemma 7B 0.125 promptemma 7B Mistral 7Bompt4 prompt2 OpenChat 3.5 OpenChat 3.5 prompt4 prompt3 rampt1 ableBeluga 7B Slower OpenChat 3.5 InternLM2.5 7B OpenChat 3.5
Ope InternLM2.5 7B InternLM2.5 7B Mississippings Pempt2 prompt2 0.100 OpenChat 3.5 prompt4 OpenChat 3.5 GPTQ prompt1 prompt2 LLaMA 3.1 8B prompt3 OWEGEN BERNES YI 34BLLaMA 3.1 8B prompt1 Or OpenChat 3.5 procompetaChat 3.5 GPTQ 0.075 prompt1 **Better Quality** 0 1000 2000 3000 4000 5000 Generation Time (seconds)