Efficiency Analysis: Quality vs. Speed 0.250 GPT-2-Large prompt2 Lower Quality 0.225 GPT-2-Medium prompt2 Yi 6B 0.200 prompti46B LLaMA Brompt3 OpenChat 3.5 GPTQ prompt4 GPT-2-SMAN PROM**REGIONA** (2018) BEINGULUE (508) 68) Quality Score (distribution_ 0.175 LLaMA 2 13B Instruct (Exp68) GPStableodilunga 7B pro**prota**pt4 InternLM2.5 7B InternLM2.5 7B LLaMA 2 13B InternLM2.5 7B Gemma 7B prompt4 prompt4 InternLM2.5 7B prompt3 PSANCELIGA 7B 3B Chat DEMINISCENSIA AMA 3 8B prohipus Hangus Balasia prohipus Hangus Balasia Gemma 7B promisemma 7B Straight Charles 5, 0106 Straight Charles 5, 0106 Straight Charles 5, 0106 Straight Charles 5, 13B Grander Charles 13B Grander 13B Grander Charles 13B Grander 13B Gran 0.125 prompt2 LLaMA 2 13B Chat Slower OpenChat 3.5 prompt2 Vicuna 13Bluggsrull fines Mistral 7B DPO InternLM2.5 7B OpenChat 3.5 7B LLaMA 3.1 8B prompt3 internLM2.5 7B prompt4 prompt3 at 3.5 GPrompt1 prompt4 prompt3 LLaMA 3.1 8B InternLM2.5 7B prompt2 0.100 OpenChat 3.5 prompt2 A STATE OF THE STA prompt4 OpenChat 3.5 GPTQ LLaMA 3.18B prompt1 prompt2 prompt3 We form 5 permes Yi 34BLLaMA 3.1 8B OpenChat 3.5 GPTQ 0.075 **Better Quality** prompt1 0 1000 2000 3000 4000 5000 Generation Time (seconds)