Efficiency Analysis: Quality vs. Speed Stable St 0.300 GEP-7-Medium GPT-2-Medium GPT-2-Medium prBfAMPt2-2-Largerompt2 prompt1 prompt3 0.275 GPT-2-Large Yi 6B prompt1Qwehr2778t1 GerAM PB prompt1 Qwen2 7B prompt4 InternLM2.5 7B OpenChar 3.5 prompt3 0.250 Quality Score (distribution_score) Promote 1335 promote 13.5 prompt3 0.225 OpenChat 3.5 GPTQ Zephyr 7B prompt4 prompt3 Zephyr 7B Nous Hermes Yi 34B prompt2 Slower Faster prompt3 LaMA 3.1 8B ZeprbymptB proprintial alsiA 3.18B prompp2ompt3 LLaMA 3.1 8B Gemma 7B prompt2 InternLM2.5 7B prompt4 prompt2 The Manager 2 7B The Manager 2 7B The Manager 2 7B To prompt 2 Objects Heer Ones Yi 34B penChat 3.5 ppromppt4 **21@phyt2**7B prompt1 0.175 OpenChat 3.5 GPTQ prompt1 OpenChat 3.5 GPTQ Nous Hermes Yi 34B prompt2 OpenChat 3.5 prompt4 0.150 OpenChat 3.5 GPTQ prompt4 Nous Heaprompt2
OpenChat 3.5
prompt2 Nous Hermes Yi 34B LLaMA 3.18B prompt4 **Better Quality** 0.125 0 1000 2000 3000 4000 5000 6000 Generation Time (seconds)