Efficiency Analysis: Quality vs. Speed

Quality Score (distribution_score)

Generation Time (seconds)

Faster

Slower

Better Quality

Lower Quality

0.300

0.275

0.250

0.225

0.200

0.175

0.150

0.125

0

1000

2000

3000

4000

5000

6000

Gemma 7B prompt3

GPT-2-Medium prompt1

GPT-2-Medium prompt2

GPT-2-Large prompt3

GPT-2-Large prompt1

Yi 6B prompt1

Qwen2 7B prompt3

Gemma 7B prompt1

Qwen2 7B prompt4

InternLM2.5 7B prompt3

OpenChat 3.5 prompt3

Zephyr 7B prompt4

OpenChat 3.5 GPTQ prompt3

Zephyr 7B prompt2

Nous Hermes Yi 34B prompt3

LLaMA 3.1 8B prompt2

Gemma 7B prompt4

InternLM2.5 7B prompt2

OpenChat 3.5 prompt2

Nous Hermes Yi 34B prompt4

Zephyr 7B prompt1

OpenChat 3.5 GPTQ prompt1

OpenChat 3.5 GPTQ prompt2

Nous Hermes Yi 34B prompt1

OpenChat 3.5 prompt4

OpenChat 3.5 GPTQ prompt4

Nous Hermes Yi 34B prompt2

OpenChat 3.5 prompt2

LLaMA 3.1 8B prompt4