Model Efficiency Ranking (Lower Score is Better)

Rank	Model	Avg Duration (s)	Avg Quality	Efficiency Score
1	Nous Hermes Yi 34B	133.1813641190529	0.09568501983494558	0.07823807598547969
2	Qwen2 7B	196.47026401758194	0.09775693808998814	0.09005944477001102
3	Mistral 7B	93.83845460414886	0.10133389732260359	0.09080356006375122
4	Zephyr 7B	121.93346518278122	0.1012497503162205	0.09316928395341202
5	StableBeluga 7B	125.74621045589447	0.12870256535591507	0.17233374086311282
6	Gemma 7B	571.3062742153803	0.12178473131462948	0.19382339076856622
7	OpenChat 3.5	1521.8950508236885	0.09942879320875118	0.2178618299018783
8	Yi 6B	214.6704027056694	0.15259372393592663	0.24917207780948336
9	OpenChat 3.5 GPTQ	1569.558363854885	0.10886776342094587	0.24938219006553422
10	LLaMA 3.1 8B	2292.9194369912148	0.09063996268870883	0.26418422414506715
11	GPT-2-Medium	112.01234618822734	0.1624626832413604	0.2679766208923115
12	GPT-2-Small	37.0994664033254	0.1681618161597093	0.2773853905102267
13	GPT-2-Large	286.69558604558307	0.16586651141200512	0.2939592776280901
14	InternLM2.5 7B	2251.6786900162697	0.12086669780282441	0.3471308764643555