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	Nous Hermes Mistral 7B DPO	160.4207947254181	0.09574007211205193	0.08092400847311215
3	Vicuna 13B	164.57771295309067	0.09719585765625105	0.08548900183483701
4	Qwen2 7B	196.47026401758194	0.09775693808998814	0.09005944477001102
5	OpenChat 3.5-0106	117.63292026519775	0.10051797462031284	0.09066942494281593
6	Mistral 7B	93.83845460414886	0.10133389732260359	0.09080356006375122
7	Zephyr 7B	121.93346518278122	0.1012497503162205	0.09316928395341202
8	LLaMA 2 13B	341.09782177209854	0.11013628151865974	0.13901946509158197
9	LLaMA 2 7B Chat	114.01400148868561	0.12228344218282661	0.15281715691436576
10	StableBeluga 7B	125.74621045589447	0.12870256535591507	0.17233374086311282
11	LLaMA 2 13B Chat	486.3273178935051	0.11863691643698421	0.17690048006345296
12	Gemma 7B	571.3062742153803	0.12178473131462948	0.19382339076856622
13	OpenChat 3.5	1521.8950508236885	0.09833961179718953	0.2147350442958944
14	LLaMA 3 8B	210.5667028427124	0.14220323465431312	0.2189625742763446
15	Yi 6B	214.6704027056694	0.15259372393592663	0.24917207780948336
16	OpenChat 3.5 GPTQ	1569.558363854885	0.10886776342094587	0.24938219006553422
17	LLaMA 2 13B Instruct (Exp68)	287.4305631518364	0.15128930772367502	0.2521797361080041
18	LLaMA 3.1 8B	2292.9194369912148	0.09063996268870883	0.26418422414506715
19	GPT-2-Medium	112.01234618822734	0.1624626832413604	0.2679766208923115
20	GPT-2-Small	37.0994664033254	0.1681618161597093	0.2773853905102267
21	GPT-2-Large	286.69558604558307	0.16586651141200512	0.2939592776280901
22	InternLM2.5 7B	2251.6786900162697	0.12086669780282441	0.3471308764643555