Model Efficiency Ranking (Lower Score is Better)

Rank	Model	Avg Duration (s)	Avg Quality	Efficiency Score
1	OpenChat 3.5	239.4489974975586	0.1019743030335099	0.1254370613184054
2	Nous Hermes Yi 34B	909.1467067718506	0.08066159226419248	0.15323313831709937
3	Mistral 7B	332.51854902505875	0.11433179478692135	0.15486635727610296
4	Nous Hermes Mistral 7B DPO	181.50977087020874	0.1454989660024516	0.19259307419567073
5	Yi 6B	735.0123144984245	0.11747783458136166	0.1981994812128665
6	LLaMA 3 8B	327.743071436882	0.16673326911272224	0.24186701608126784
7	LLaMA 2 13B Chat	441.7801613807678	0.16097281540314226	0.24304326750390012
8	Zephyr 7B	1167.365066230297	0.12325289220778995	0.24874532200659427
9	InternLM2.5 7B	695.2830808758736	0.15663578294842323	0.2597908841343096
10	OpenChat 3.5 GPTQ	1788.4899263381958	0.10699097787158313	0.2803748292379145
11	Gemma 7B	973.2726464748382	0.15589085954326698	0.2848502004213457
12	OpenChat 3.5-0106	1459.9080494244893	0.13845345999063127	0.30179297037296443
13	Qwen2 7B	553.1936149597168	0.19911580497091902	0.31724146630452743
14	LLaMA 2 13B	732.658384501934	0.18974622753289724	0.3185850382626184
15	GPT-2-Medium	555.6102035840353	0.20878933774652386	0.33361421609589853
16	GPT-2-Large	735.576487382253	0.20807331442482316	0.3494471032803275
17	GPT-2-Small	208.61146799723306	0.2584541019224604	0.38366748803337086
18	LLaMA 2 7B Chat	1139.6794353485107	0.21467058340421863	0.39869221835780044
19	LLaMA 3.1 8B	1875.7036468982697	0.17329151164868106	0.39927527719463163
20	StableBeluga 7B	1415.43996489048	0.20969239839642806	0.4164757657586995