Model Efficiency Ranking (Lower Score is Better)

Rank	Model	Avg Duration (s)	Avg Quality	Efficiency Score
1	Nous Hermes Yi 34B	909.7884995937347	0.17510808043858112	0.2005198045986366
2	OpenChat 3.5	367.83977446556094	0.1986750889397208	0.22100067058580466
3	LLaMA 3.1 8B	829.9822397828102	0.18547342708834325	0.22322469082016594
4	InternLM2.5 7B	328.18671170870465	0.212110071086856	0.2559397577766496
5	Yi 6B	225.19416411717734	0.23017548889431239	0.2986297800473604
6	Zephyr 7B	1320.614428281784	0.20431958520716959	0.31942458195846585
7	Qwen2 7B	859.1666157245636	0.22491595127408587	0.3383766618156767
8	Gemma 7B	1011.7931925058365	0.24269184221883675	0.40232163736794535
9	GPT-2-Large	424.71745403607684	0.27484749890762844	0.44342967826399066
10	OpenChat 3.5 GPTQ	4268.031990170479	0.17277962317357595	0.4839733761705878
11	StableBeluga 7B	79.46426335970561	0.30130028940173087	0.4891426848927655
12	GPT-2-Small	110.13617213567098	0.30156175510698285	0.4925389183625007
13	GPT-2-Medium	922.2149715423584	0.27877112181066627	0.49761029494282544