Model	Auto-J Rating	GPT-4 Win-rate	Rank GPT-4	Rank Auto-J	Δ
XwinLM 70b V0.1	5.694	95.57	1	1	0
LLaMA2 Chat 70B	5.678	92.66	2	2	0
XwinLM 13b V0.1	5.647	91.76	3	3	0
OpenChat V3.1 13B	5.532	89.49	4	8	4
WizardLM 13B V1.2	5.547	89.17	5	6	1
Vicuna 33B v1.3	5.570	88.99	6	5	-1
Humpback LLaMa2 70B	5.498	87.94	7	11	4
XwinLM 7b V0.1	5.584	87.83	8	4	-4
OpenBudddy-LLaMA2-70B-v10.1	5.448	87.67	9	14	5
OpenChat V2-W 13B	5.533	87.13	10	7	-3
OpenBuddy-LLaMA-65B-v8	5.458	86.53	11	13	2
WizardLM 13B V1.1	5.497	86.32	12	12	0
OpenChat V2 13B	5.519	84.97	13	9	-4
Humpback LLaMa 65B	5.379	83.71	14	19	5
Vicuna 13B v1.3	5.388	82.11	15	18	3
OpenBuddy-LLaMA-30B-v7.1	5.391	81.55	16	17	1
LLaMA2 Chat 13B	5.518	81.09	17	10	-7
OpenChat-13B	5.437	80.87	18	15	-3
OpenBuddy-Falcon-40B-v9	5.373	80.70	19	20	1
UltraLM 13B	5.342	80.64	20	22	2
OpenChat8192-13B	5.429	79.54	21	16	-5
OpenCoderPlus-15B	5.357	78.70	22	21	-1
OpenBudddy-LLaMA2-13B-v11.1	5.340	77.49	23	23	0
Vicuna 7B v1.3	5.332	76.84	24	25	1
WizardLM 13B	5.247	75.31	25	32	7
JinaChat	5.319	74.13	26	26	0
airoboros 65B	5.318	73.91	27	27	0
airoboros 33B	5.289	73.29	28	30	2
Guanaco 65B	5.313	71.80	29	29	0
LLaMA2 Chat 7B	5.334	71.37	30	24	-6
Vicuna 13B	5.314	70.43	31	28	-3
OpenBuddy-Falcon-7b-v6	5.214	70.36	32	34	2
Baize-v2 13B	5.165	66.96	33	38	5
LLaMA 33B OASST RLHF	5.173	66.52	34	37	3
Minotaur 13B	5.210	66.02	35	36	1
Guanaco 33B	5.212	65.96	36	35	-1
Nous Hermes 13B	5.271	65.47	37	31	-6
Vicuna 7B	5.237	64.41	38	33	-5
Baize-v2 7B	5.083	63.85	39	39	0
LLaMA 33B OASST SFT	4.985	54.97	40	41	1
Guanaco 13B	5.027	52.61	41	40	-1
ChatGLM2-6B	4.846	47.13	42	46	4
Guanaco 7B	4.943	46.58	43	43	0
Falcon 40B Instruct	4.934	45.71	44	44	0
Alpaca Farm PPO Sim (GPT-4) 7B	4.978	44.10	45	42	-3
Pythia 12B SFT	4.809	41.86	46	47	1
Alpaca Farm PPO Human 7B	4.907	41.24	47	45	-2
Cohere Chat	4.524	29.57	48	51	3
Cohere	4.522	28.39	49	52	3
Alpaca 7B	4.658	26.46	50	48	-2
Pythia 12B OASST SFT	4.620	25.96	51	49	-2
Falcon 7B Instruct	4.537	23.60	52	50	-2
Baichuan-13B-Chat	4.291	21.80	53	53	0

Table 24: Values and ranking by Auto-J and GPT-4 for open-source LLMs on AlpacaEval. Value of AUTO-J is the model's average rating on AlpacaEval dataset assigned by AUTO-J in single-response evaluation protocol, and value of GPT-4 is the model's win-rate against Davinci003 determined by GPT-4 on AlpacaEval dataset. $\Delta = \mathsf{Rank}_{\mathsf{Auto\text{-}J}} - \mathsf{Rank}_{\mathsf{GPT\text{-}4}}$