Method	GPT-3.5 Turbo	Llama-8B	Mistral-Nemo	Llama-405B	Mistral-Large-2	GPT-4 Turbo	GPT-4.1	GPT-4o	DeepSeek-V3	o3-mini (R)	o3 (R)	o4-mini (R)	DeepSeek-R1 (R)
MAD	0.194	0.586	0.681	0.946	0.914	0.965	0.955	0.948	0.975	0.975	0.940	0.961	0.954
MAB	0.313	0.545	0.714	0.950	0.938	0.955	0.960	0.964	0.970	0.984	0.956	0.957	0.960
SC-5	0.517	0.535	0.613	0.925	0.912	0.950	0.941	0.961	0.985	0.985	0.970	0.980	0.990
SC-3	0.515	0.524	0.613	0.925	0.912	0.945	0.941	0.961	0.974	0.985	0.975	0.980	0.990
CoT-5	0.521	0.524	0.632	0.925	0.911	0.929	0.918	0.957	0.970	0.979	0.975	0.966	0.980
CoT-3	0.389	0.309	0.634	0.938	0.899	0.921	0.940	0.952	0.985	0.980	0.975	0.971	0.985
CoT-1	0.234	0.340	0.302	0.847	0.903	0.945	0.935	0.954	0.970	0.984	0.970	0.980	0.990
I/O	0.464	0.431	0.598	0.802	0.808	0.928	0.956	0.951	0.975	0.985	0.956	0.952	0.975