HELPFULNESS EVALUATION

Dataset	LLaVA Eval				LLaVA Bench			
Model	Conversation	Description	Reasoning	Average	Relevance	Accuracy	Level of detail	Helpfulness
BLIP-2	66.08	31.33	22.00	39.80	25.00	16.00	16.00	17.67
InstructBLIP	74.08	61.67	82.17	72.64	34.00	21.00	19.67	22.67
LLaVA	65.17	42.17	61.50	56.28	31.83	19.83	18.67	20.83
LLaVA-HF	69.74	60.87	**85.33**	71.98	34.33	18.50	17.67	23.50
mPLUG	66.08	44.17	75.83	62.03	35.17	**20.33**	16.33	20.33
miniGPT4	54.92	51.50	74.67	60.36	32.45	**20.33**	20.17	24.17
DRESS	**77.67**	**62.17**	84.27	**74.70**	**37.18**	20.12	**21.87**	**26.45**