	GPT-4 judgments						Human judgments					
Baseline judgment	SelFee			Vicuna			SelFee			Vicuna		
	Win	Tie	Lose	Win	Tie	Lose	Win	Tie	Lose	Win	Tie	Lose
Summarization	12	0	0	9	0	3	10	1	1	11	1	0
Exam Questions	11	1	0	7	1	4	8	2	2	7	2	3
Code	20	0	0	15	0	5	15	2	3	11	5	4
Rewriting	18	0	2	18	0	2	14	5	1	14	4	2
Creative Writing	33	0	3	29	0	7	26	7	3	24	9	3
Functional Writing	37	0	3	28	0	12	37	2	1	32	6	2
General Communication	47	0	1	39	0	9	43	2	3	39	6	3
NLP Tasks	40	0	4	35	0	9	29	7	8	22	15	7
Overall	218	1	13	180	1	51	182	28	22	160	48	24
Baseline judgment	L2Chat			ChatGPT			L2Chat			ChatGPT		
	Win	Tie	Lose	Win	Tie	Lose	Win	Tie	Lose	Win	Tie	Lose
Summarization	10	0	2	12	0	0	10	1	1	10	1	1
Exam Questions	11	0	1	10	0	2	6	3	3	6	1	5
Code	18	0	2	16	0	4	11	5	4	8	6	6
Rewriting	17	1	2	14	0	6	10	7	3	9	8	3
Creative Writing	30	0	6	26	2	8	13	14	9	16	15	5
Functional Writing	32	0	8	22	2	16	23	13	4	23	14	3
General Communication	41	0	7	36	0	12	28	15	5	29	10	9
NLP Tasks	37	1	6	35	1	8	13	22	9	16	17	11
Overall	196	2	34	171	5	56	114	80	38	117	72	43
Baseline judgment	WizardLM			GPT-4			WizardLM			GPT-4		
	Win	Tie	Lose	Win	Tie	Lose	Win	Tie	Lose	Win	Tie	Lose
Summarization	10	0	2	8	0	4	11	1	0	6	2	4
Exam Questions	11	1	0	3	0	9	6	3	3	2	2	8
Code	17	0	3	12	0	8	9	7	4	6	5	9
Rewriting	15	1	4	12	0	8	12	6	2	12	3	5
Creative Writing	31	1	4	23	0	13	17	17	2	11	9	16
Functional Writing	30	1	9	17	1	22	24	12	4	27	6	7
General Communication	38	1	9	28	0	20	32	14	2	35	3	10
NLP Tasks	35	2	7	23	1	20	19	14	11	8	11	25
Overall	187	7	38	126	2	104	130	74	28	107	41	84

Table 23: Detailed comparison results between critiques generated by AUTO-J and baselines for single-response evaluation. Results on left side are GPT-4 judgments, and results on right side are human judgments. Vicuna, L2Chat, and WizardLM respectively stand for Vicuna-13B-v1.5, LLaMA-2-Chat-13B, and WizardLM-13B-v1.2.