Label Distribution (Label, # of Samples)					
Win	1594	Lose	1596	Tie	246
Source Dataset Distribution (Source, # of Samples)					
Chatbot Arena Conversations	2801	OpenAI Summary	100	OpenAI WebGPT	45
PKU-SafeRLHF	158	Stanford SHP	81	Synthetic GPT-J	251
Scenario Distribution (Name, # of Samples)					
ranking	100	open_question	100	text_correction	18
recommendation	100	post_summarization	100	writing_product_description	16
creative_writing	100	writing_song_lyrics	98	language_polishing	15
planning	100	functional_writing	94	code_to_code_translation	15
brainstorming	100	writing_cooking_recipe	88	writing_legal_document	13
exam_question_without_math	100	code_correction_rewriting	86	writing_blog_post	13
roleplay	100	writing_personal_essay	84	title_generation	12
text_summarization	100	analyzing_general	67	writing_social_media_post	12
asking_how_to_question	100	explaining_code	59	reading_comprehension	11
chitchat	100	information_extraction	51	writing_technical_document	10
verifying_fact	100	writing_email	51	text_simplification	10
value_judgment	100	writing_job_application	46	keywords_extraction	6
code_generation	100	classification_identification	44	writing_scientific_paper	5
text_to_text_translation	100	writing_presentation_script	42	writing_marketing_materials	4
math_reasoning	100	exam_question_with_math	41	topic_modeling	3
question_generation	100	data_analysis	39	writing_news_article	3
counterfactual	100	instructional_rewriting	30	note_summarization	2
seeking_advice	100	paraphrasing	27	code_simplification	1
explaining_general	100	writing_advertisement	20	others	100

Table 21: Statistics for pairwise training data: the distribution of labels, source datasets, and scenarios.

Score Distribution (Score, # of Samples)					
1	29	2	137	3	178
4	210	5 (5.5)	131	6 (6.5)	241
7	27	8	4	10	3
Scenario Distribution (Name, # of Samples)					
code_generation	24	explaining_code	18	writing_technical_document	15
explaining_general	23	functional_writing	18	text_simplification	15
open_question	23	writing_song_lyrics	18	language_polishing	15
seeking_advice	23	ranking	18	code_to_code_translation	15
math_reasoning	22	planning	17	writing_blog_post	15
chitchat	21	classification_identification	17	reading_comprehension	14
value_judgment	21	exam_question_with_math	17	topic_modeling	14
brainstorming	21	writing_cooking_recipe	17	writing_advertisement	14
creative_writing	20	writing_email	17	title_generation	14
roleplay	20	information_extraction	17	keywords_extraction	14
verifying_fact	20	paraphrasing	17	writing_legal_document	14
counterfactual	19	code_correction_rewriting	17	writing_news_article	14
asking_how_to_question	19	data_analysis	16	writing_social_media_post	14
exam_question_without_math	19	writing_product_description	16	code_simplification	12
text_summarization	19	instructional_rewriting	16	writing_scientific_paper	12
recommendation	18	writing_presentation_script	16	writing_marketing_materials	8
question_generation	18	analyzing_general	16	note_summarization	4
text_to_text_translation	18	writing_job_application	16	writing_biography	4
writing_personal_essay	18	text_correction	16	others	27

Table 22: Statistics for single training data: the distribution of GPT-4 ratings, and scenarios.