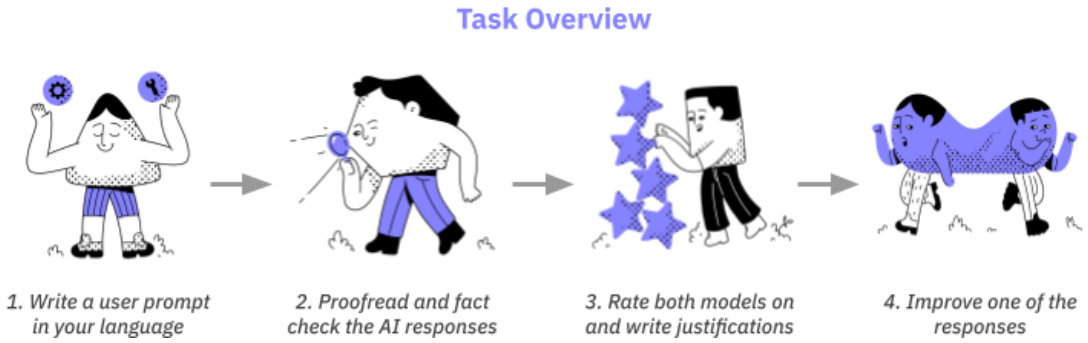
Winter Wonderland RLHF | Do's and Don'ts

Instructions | Winter Wonderland RLHF

✅ High-Level Tips



	Do!	Don't
Prompt	• Want to stick to the assigned category for the prompt • Want the prompt to be specific and answerable • Adhere to the desired complexity level (ie # of constraints)	• Choose a different category, you must stick to your category • Have proofreading errors or poor writing • Don't pick something out of your expertise or that is too hard to assess in terms of the response
Response Ratings	• Ensure that your dimensional ratings are objective and at least 9/10 of people would agree with you • Fact-check every claim the model makes. Models often make up information! • Pay attention to language fluency! The model will often use incorrect	• Not fact check!!! • Assume the model is fluent in your language • Don't select responses where theres a ton of complex/hard fact checking involved • "Make up an error" → be intellectually honest in the way you rate the model

	grammar, awkward phrasing or literal translations! • Spend time and be thorough on the ratings → make sure that you write your justifications in a way that helps the reviewer understand the issues you saw • Provide evidence for your ratings in the dimension-specific justification boxes	
Rewrite	• Remove pleasantries • Make sure the tone of your response is warm • Double check any factual information you are adding to the response	• Leave issues that you identified in your ratings unresolved (fix everything you identified)!

Side-by-side Rating	• Write a thorough justification explaining:
	  ○ Your verdict
	  ○ Why the response is better
	  ○ Specifically what it does better on
	  ○ When you write your justification make sure to explain your thinking in detail (even if you have already explained it in the dimension-specific justification)
	  ○ Your final justification should include everything you have mentioned in the dimension-specific justifications
	• TRIPLE CHECK YOU'RE REFERENCING THE RIGHT RESPONSE (CHECK THAT RESPONSE 1 AND RESPONSE 2) ARE WHAT THEY ARE.
	• Make sure your side-by-side rating
		• Be generic in your justification
		• Make your side-by-side rating inconsistent with your dimension justifications (ie if one response is rated higher on most dimensions you should not prefer the other response on your side-by-side rating)

	aligns with your dimensional ratings (ie if one response is rated higher on most dimensions, it should generally be preferred on the side-by-side rating)	