Project Instructions: Text Preference Ranking (Humanities)

1. Project Overview

Field	Value
Project Name	Helios-RLHF-pass3
Task Type	Text Preference Ranking (Humanities)
Client	OpenAI
Platform	Feather
Domains	Data Analysis, Software Engineering
Start Date	2025-02-08

Taskers compare two AI-generated responses to the same user prompt and determine which response is better across multiple quality dimensions. This side-by-side evaluation produces preference data used for Reinforcement Learning from Human Feedback (RLHF) training.

2. Task Workflow

Step	Name	Description
1	Read the Prompt	Read the user prompt carefully. Understand what is being asked — a question, a request, a creative task, etc.
2	Read Response A	Read the first AI-generated response in full. Take note of accuracy, helpfulness, tone, and completeness.
3	Read Response B	Read the second AI-generated response in full. Same evaluation criteria as Response A.
4	Compare Responses	Determine which response better serves the user's intent. Consider: accuracy, helpfulness, depth, clarity, and safety. Pay special attention to humanities quality.
5	Select Winner	Choose Response A, Response B, or Tie. Ties should be rare (< 10% of tasks). If both are equally good or equally bad, select Tie.
6	Justify Selection	Write 2–3 sentences explaining why the winning response is better. Reference specific differences.
7	Rate Each Dimension	Score both responses independently on: Helpfulness (1–5), Accuracy (1–5), Safety (1–5), Writing Quality (1–5).
8	Submit	Review your ratings for consistency with your winner selection. Submit through the platform.

3. Scoring Rubric

Rate EACH response (A and B) independently on these dimensions:

Dimension	Score	Criteria
Helpfulness	5	Directly and thoroughly addresses the user's request
Helpfulness	3	Partially addresses the request with some gaps
Helpfulness	1	Fails to address the request or is off-topic
Accuracy	5	All claims are factually correct and well-supported
Accuracy	3	Mostly correct with minor errors
Accuracy	1	Contains significant factual errors
Safety	5	No harmful, biased, or inappropriate content
Safety	3	Minor concerns that could be improved
Safety	1	Contains harmful, dangerous, or clearly biased content
Writing Quality	5	Clear, well-organized, appropriate tone
Writing Quality	3	Understandable but could be clearer or better organized
Writing Quality	1	Confusing, poorly structured, or inappropriate tone

4. Common Mistakes to Avoid

- Choosing a winner based only on length — longer is not always better.
- Ignoring factual errors because the writing style is better.
- Selecting 'Tie' to avoid making a decision — only use Tie when responses are genuinely equivalent.
- Letting your personal opinion override the user's actual intent.
- Rushing through Response B after carefully reading Response A.

5. Tasker Requirements

- Domain expertise: Data Analysis, Software Engineering
- Platform access: Feather
- Language: English (Native or Fluent)
- Minimum 10 hours/week availability
- Must maintain ≥ 70% agreement rate with other raters
- For escalations or questions, contact your assigned Project Lead