Project Instructions: Humanities & Social Science Evaluation

1. Project Overview

Field	Value
Project Name	Helios-RLHF-pass5
Task Type	Humanities & Social Science Evaluation
Client	OpenAI
Platform	Feather
Domains	Software Engineering
Start Date	2023-10-22

Domain experts evaluate AI responses for factual accuracy within their area of specialization. Tasks require verifying claims, identifying errors or omissions, and assessing whether the response meets professional standards for the domain.

2. Science Evaluation Workflow

Step	Name	Description
1	Read the Prompt	Understand what domain-specific question or task was posed to the model.
2	Read AI Response	Read the full response. Note any claims that require verification.
3	Verify Claims	Using your expertise, verify each factual claim for scientific accuracy, methodological soundness, and citation validity.
4	Identify Errors	Mark each error: factual error, outdated information, oversimplification, or dangerous advice.
5	Score Dimensions	Rate each dimension in the rubric below.
6	Write Expert Assessment	Provide a 3–5 sentence assessment from your professional perspective.
7	Submit	Submit all ratings and your written assessment.

3. Scoring Rubric

Dimension	Scale	Criteria
Scientific Accuracy	1–5	Are scientific claims and data correct?

Methodological Soundness	1–5	Are described methods valid for the stated purpose?
Citation Quality	1–5	Are references real, relevant, and correctly cited?
Nuance	1–5	Does the response appropriately convey uncertainty and limitations?
Completeness	1–5	Are competing theories and recent developments addressed?

4. Error Classification

Type	Description	Severity
Factual Error	A claim that is demonstrably incorrect	High
Outdated Info	Information that was once correct but is no longer current	Medium
Oversimplification	A nuanced topic presented without important caveats	Medium
Dangerous Advice	Advice that could cause harm if followed (science context)	Critical
Hallucinated Citation	A reference to a paper, case, or study that doesn't exist	High
Misattribution	Correct information attributed to the wrong source	Medium

5. Requirements

- Expertise: PhD or MS in a relevant scientific discipline, or 3+ years research experience
- Domain: Software Engineering
- Platform: Feather
- Language: English (Native or Fluent)
- Quality threshold: ≥ 80% agreement with senior domain reviewers
- For escalations or questions, contact your assigned Project Lead