Project Instructions: Reasoning & Logic Assessment

1. Project Overview

Field	Value
Project Name	Claude-HumanitiesEval-2024
Task Type	Reasoning & Logic Assessment
Client	Anthropic
Platform	Airtable
Domains	Humanities, Physical Sciences, Social Sciences
Start Date	2024-11-06

Taskers evaluate AI model outputs on mathematical, logical, and analytical reasoning tasks. Each task presents a problem, the model's step-by-step reasoning chain, and its final answer. Taskers verify each step and identify where reasoning breaks down.

2. Task Workflow

Step	Name	Description
1	Read the Problem	Understand the mathematical, logical, or analytical problem presented.
2	Solve Independently	Before reading the AI's answer, solve the problem yourself (or at least outline the approach). This prevents anchoring bias.
3	Read Chain of Thought	Read the model's step-by-step reasoning. Check each step for logical validity.
4	Identify First Error	If there's an error, mark the FIRST step where reasoning goes wrong. All subsequent steps are tainted.
5	Verify Final Answer	Is the final answer correct, regardless of the reasoning path?
6	Score Dimensions	Rate: Reasoning Quality (1–5), Answer Correctness (Binary), Explanation Clarity (1–5).
7	Classify Error Type	If wrong: Arithmetic Error, Logic Error, Misunderstanding, Incomplete Analysis, or Correct Reasoning/Wrong Answer.
8	Submit	Submit scores, error classification, and the step number of first error (if applicable).

3. Error Type Classification

Type	Description	Example

Arithmetic	Calculator-level mistake (2+2=5, wrong multiplication, sign error)	Common in multi-step calculations
Logic	Invalid inference or logical fallacy (affirming consequent, false equivalence)	Model says A→B therefore B→A
Misunderstanding	Model misinterprets the problem statement	Solves a different problem than asked
Incomplete	Reasoning stops short or misses cases	Proves for n=1 but doesn't complete induction
Hallucinated Step	Model invents a fact or theorem that doesn't exist	Cites 'theorem' that is not real

4. Reasoning Quality Rubric (1–5)

Score	Criteria
5	Flawless reasoning chain. Every step is valid and clearly explained.
4	Sound reasoning with minor presentation issues. All steps valid.
3	Mostly correct reasoning with 1 non-critical error or unclear step.
2	Multiple reasoning errors or a critical logical flaw.
1	Fundamentally flawed reasoning or no coherent chain of thought.

5. Requirements

- Strong quantitative background (STEM degree or equivalent experience)
- Platform: Airtable
- Must be comfortable with: algebra, calculus, probability, logic, and basic proof techniques
- Quality threshold: ≥ 80% agreement with expert solutions
- For escalations or questions, contact your assigned Project Lead