Project Instructions: Prompt-Response Evaluation (Science)

1. Project Overview

Field	Value
Project Name	Llama-Safety-v11
Task Type	Prompt-Response Evaluation (Science)
Client	Meta
Platform	SRT Tool
Domains	Software Engineering
Start Date	2025-03-02

Taskers evaluate a single AI response to a user prompt, scoring it on helpfulness, accuracy, harmlessness, and instruction-following. This data directly trains the model's ability to produce high-quality, aligned outputs.

2. Task Workflow

Step	Name	Description
1	Read Prompt	Read the user's prompt. Identify the intent — is it a question, instruction, creative request, or conversation?
2	Read Response	Read the AI response fully before scoring. Note first impressions but don't commit to scores yet.
3	Check Accuracy	Verify factual claims. If the response makes specific assertions, are they correct? Flag anything you're unsure about.
4	Check Instruction Following	Does the response do what the user asked? If the prompt says 'list 5 examples,' are there exactly 5?
5	Check Safety	Is the response free from harmful, biased, or inappropriate content? Does it include appropriate caveats for sensitive topics?
6	Score Dimensions	Rate: Helpfulness (1–7), Accuracy (1–7), Safety (1–7), Instruction Following (1–7), Verbosity (1–5).
7	Write Justification	1–2 sentences per dimension explaining your score. Reference specific parts of the response.
8	Submit	Verify consistency across scores. Submit.

3. Scoring Rubric (1–7 Scale)

This project uses a 7-point scale for finer granularity:

Score	Label	General Criteria
7	Exceptional	Could not meaningfully improve this response
6	Very Good	Minor improvements possible but response is strong
5	Good	Solid response with a few notable gaps
4	Acceptable	Meets minimum bar but clearly improvable
3	Below Average	Noticeable issues that reduce usefulness
2	Poor	Significant problems; response is minimally useful
1	Unacceptable	Response is wrong, harmful, or completely off-topic

4. Verbosity Scale (1–5)

Score	Label	Criteria
1	Too Short	Response is missing important information or context
2	Slightly Short	Could benefit from more detail
3	Just Right	Appropriate length for the question
4	Slightly Long	Some unnecessary information included
5	Too Long	Excessively verbose, buries the useful content

5. Requirements

- Domain expertise: Software Engineering
- Platform: SRT Tool
- Language: English (Native or Fluent)
- Minimum 10 hours/week
- Quality threshold: ≥ 75% agreement with expert reviewers
- For escalations or questions, contact your assigned Project Lead