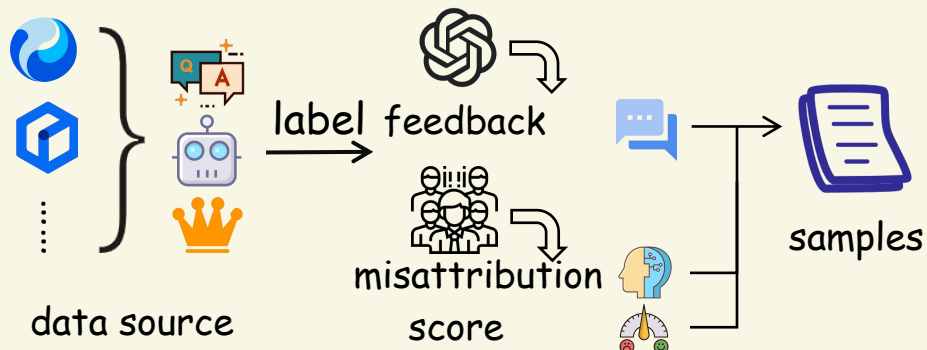



## I Data Construction




### Instruction:

 You are a helpful judge model .....


### Qusetion:

 Who was the first person to walk on the moon?

### Model Answer:

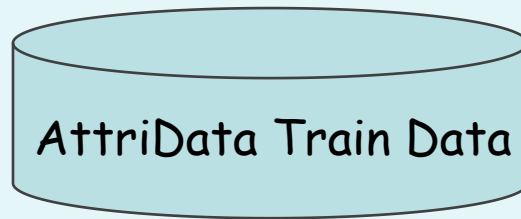
 The first person to walk on the moon was Charles Lindbergh in 1951, during the LunarPioneer mission.

### Reference:

 Neil Armstrong was the first person to walk on the moon in 1969 during the Apollo 11 mission.



review



## II Supervised Fine-tuning

input

MisAttributionLLM

## III Inference



- **Feedback:** The model answered incorrectly. Neil Armstrong was the first person to walk on the moon, suggesting a potential hallucination problem.



- **Misattribution:** Knowledge Ability-Hallucination



- **Score:** 1