Rule Annotation

Rules
- Be obedient.
- Avoid risky for human.
- Do not be sycophantic.
- Remember ...
- .....

Tags
- Direction
- Stage
- Domain
- Family

Specification Generation

Constraints
- Direction Balance
- Stage Coverage
- Theme Coverage
- Family Diversity

Priority
- Stage
- Direction
- Exception
- Specificity

Multi-Agent Adversarial Data Synthesis

1 Planning
- Model Spec
- Top-K Examples
- Model Spec
- Planner

Plan

2 Adversarial Interaction
Attack prompt

Role Switching (New Seed)

Attack prompt

Experience Bank
- Successful Entry 1
- Successful Entry 2
- Successful Entry 3
- Successful Entry 4
- .....

3 Evaluation
Success?

Safe?

Helpful?

User Input

F1 F2 F3 F4 F5 F6 F7

D1 D2 D3

Summary Agent

○ Fast Agent ○ Deep Agent

User Input [BEGIN]

Teams 1 — Report — Assign

Center

Team 2 — Report — Assign

t=2, 3... 5

User [BEGIN] → State agent — Target, True value → Agent₁ — Target, Reported value → Agent₂

State agent ← Agent₁ and Agent₂'s Choice — Agent₁ ← Agent₂'s Choice — Agent₂

t=1, [BEGIN]