INDUCTION STEP 1

Feedback Grounding

Agent **Trajectory** + Human **Feedback**
→ LLM → some **behavior** in the trajectory corresponding feedback **aspect**

M pairs of trajectory and feedback

Aspects

INDUCTION STEP 2

Behavior Clustering

Aspects from all **M** trajectories → LLM → **Metric Description**, **Good Behaviors**, **Bad Behaviors** xN

AutoLibra **Metrics**

EVALUATION STEP 1

Evaluate Agents

+ **Metric Description**, **Good Behaviors**, **Bad Behaviors** xN → LLM → xN

Evaluate each trajectory with induced **metrics**

Traits

EVALUATION STEP 2

Meta Evaluation

Aspects ←LLM→ **Traits** Match xM → aggregation → Coverage / Redundancy