Stanford University

Caltech

Carnegie Mellon University

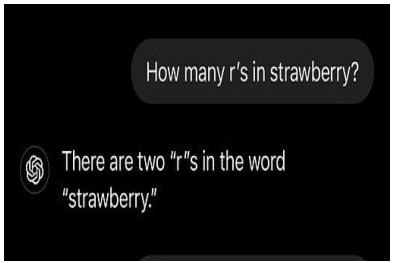
上海人工智能实验室
Shanghai Artificial Intelligence Laboratory

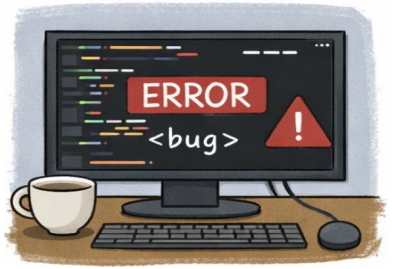
From Reasoning Failures to Reliable Intelligence

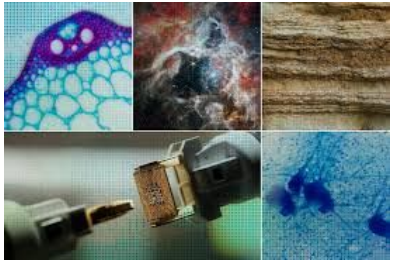
Understanding, Evaluating, and Building Better LLM Reasoning Systems

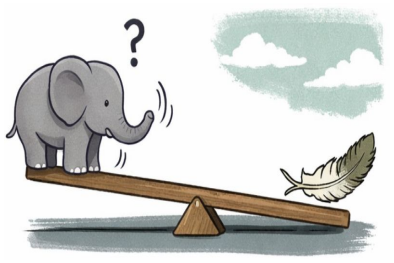
Peiyang Song

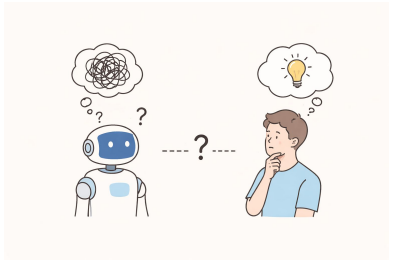
Why study reasoning failures, at all?





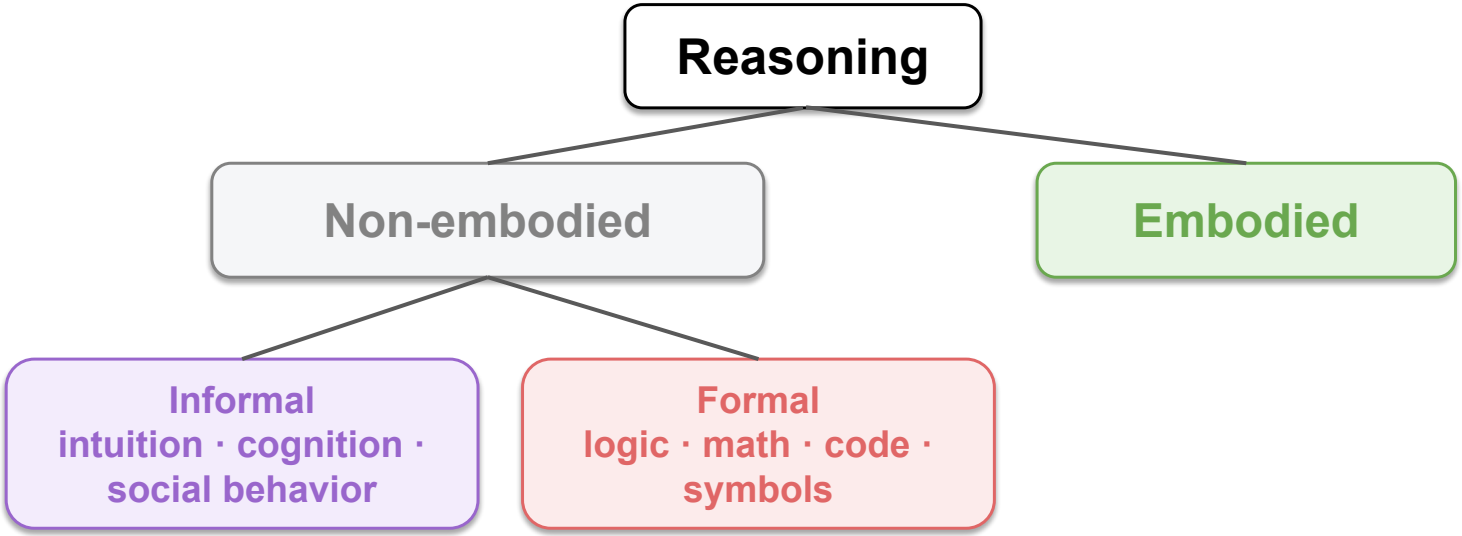






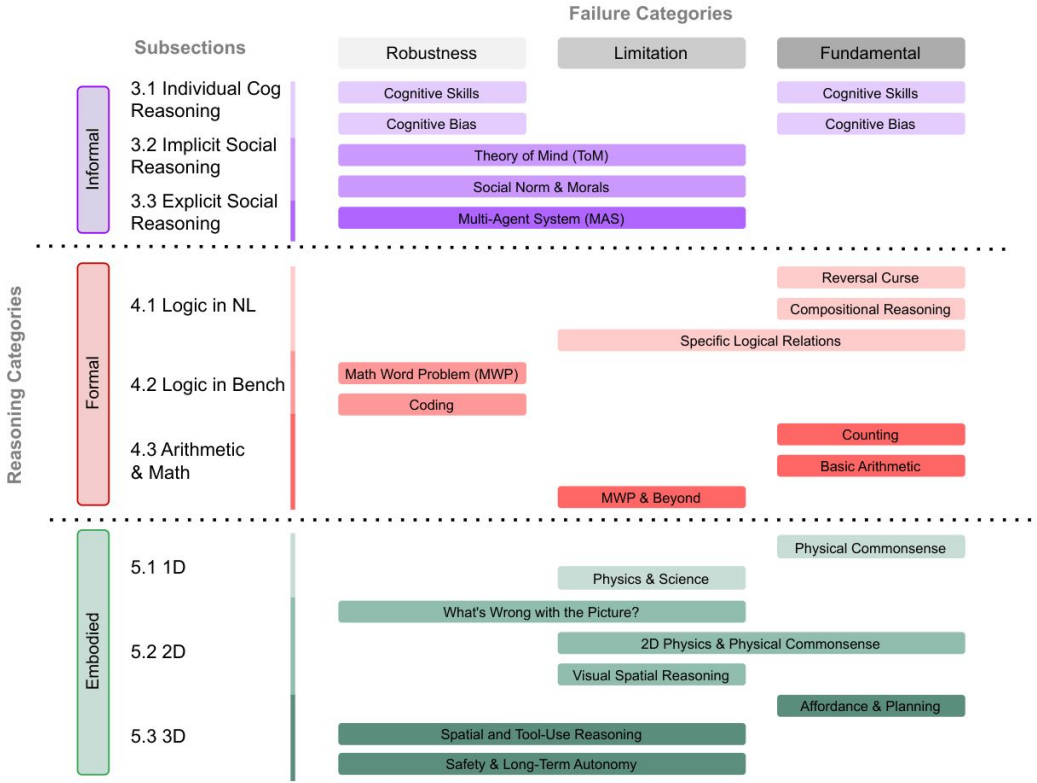
Failure analysis is not peripheral; it is central to understanding capability.

What kinds of reasoning are we talking about?



Different forms of reasoning may fail differently, and offer different routes to improvement.

A two-axis taxonomy: reasoning x failure



Reasoning categories

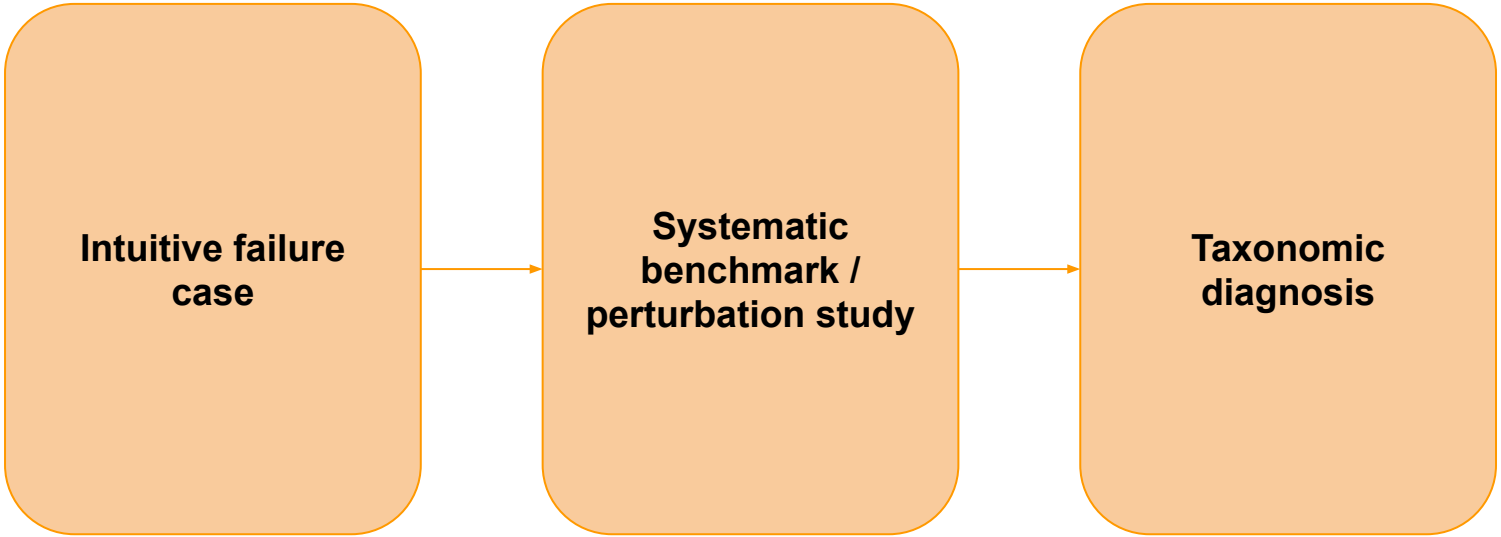
- Non-embodied
  a. Informal reasoning
  b. Formal reasoning
- Embodied reasoning

Failure categories

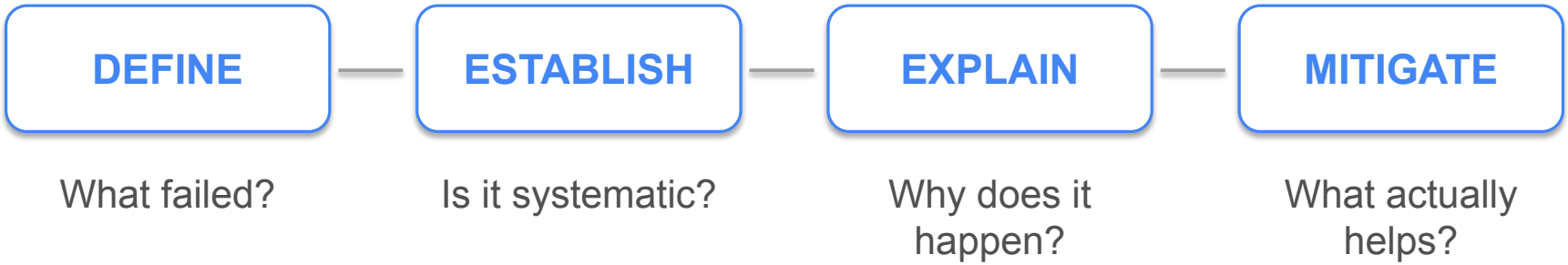
- Fundamental failures
- Application-specific limitations
- Robustness issues

From failure anecdotes to failure science

Failure research usually starts from a striking example, but should end with structured diagnosis.



What should we learn from each failure?



A catalogue records failures; a science of failures explains and predicts them.

Informal reasoning: individual cognition

Missing cognitive skills

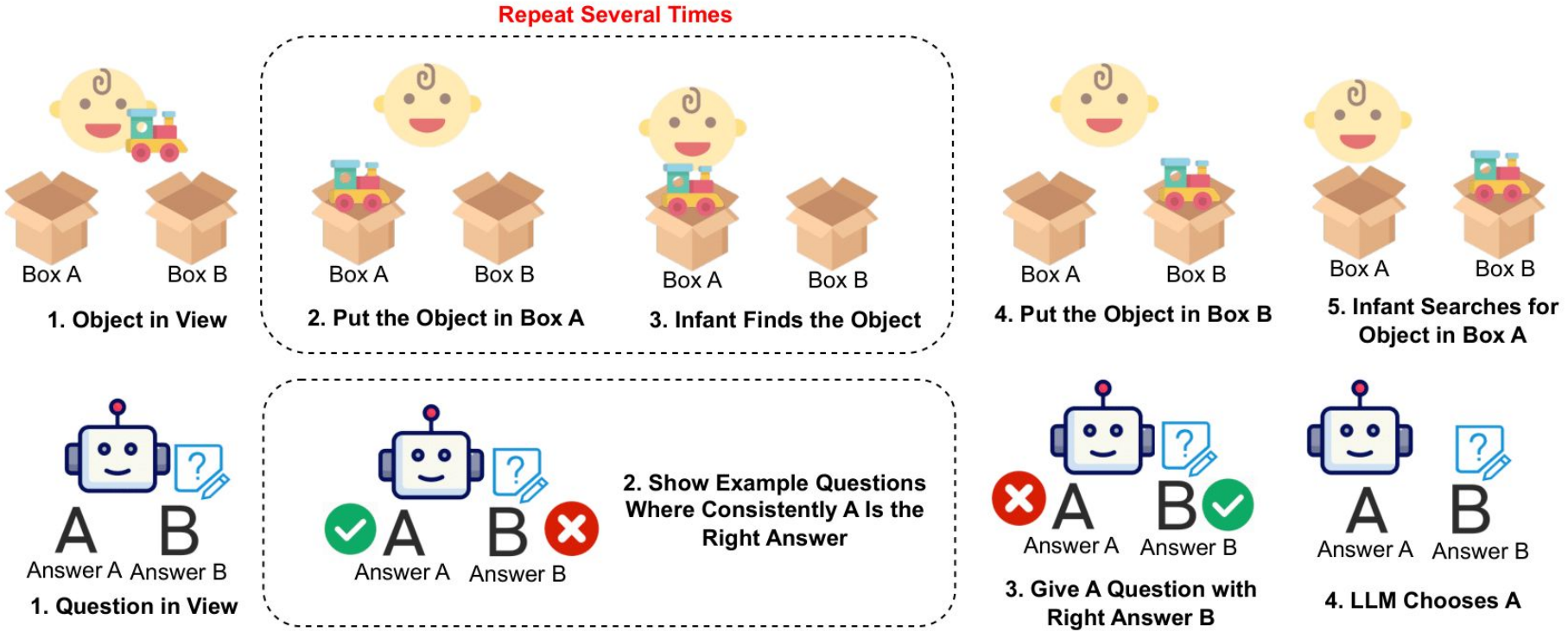
- working memory
- inhibitory control
- cognitive flexibility
- abstract reasoning

Cognitive biases

- confirmation bias
- order bias
- anchoring bias
- framing effects

LLMs fail both because they lack some cognitive skills, and because they inherit some human-like biases.

A concrete failure example: inhibitory control



Informal reasoning: social reasoning

Implicit social reasoning

- Theory of Mind
- emotion / perspective / false belief
- social norms and moral values

Explicit social reasoning

- multi-agent planning
- communication and coordination
- robustness and verification

Saying the socially appropriate thing is not the same as behaving consistently in social contexts.

Formal reasoning: logic in natural language

Representative failures

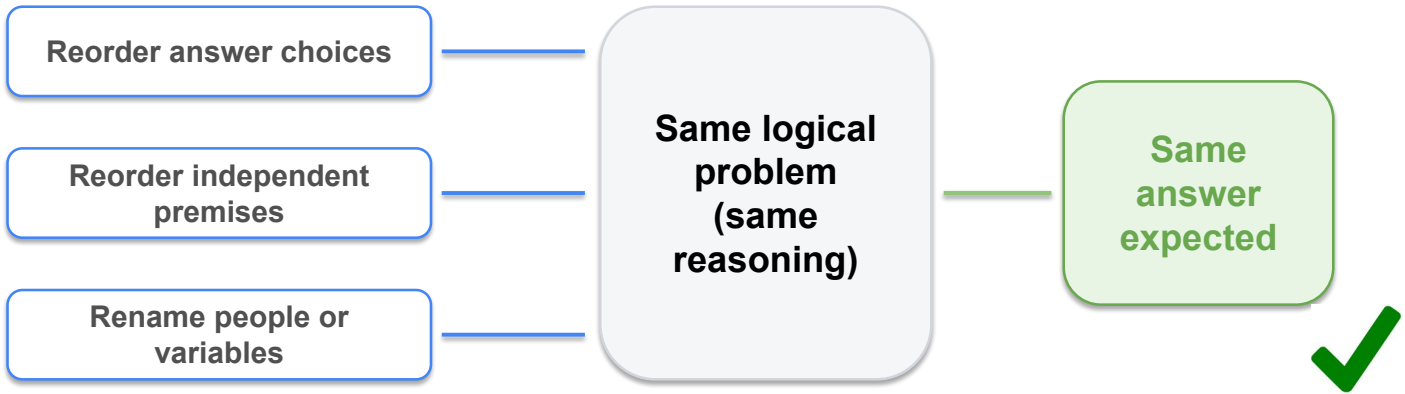
- Reversal curse
- Compositional reasoning breakdown
- Weaknesses on specific logical relations

The model may know pieces of logic without reliably representing logical structure.

Formal reasoning: benchmark robustness

Strong static benchmark scores can hide weak reasoning robustness.

- If the answer changes when irrelevant details change, benchmark success may overstate competence.
- Key method: **logic-preserving perturbations**



Formal reasoning: arithmetic and mathematics

Fundamental failures

- counting remains brittle
- arithmetic degrades as operands grow
- pattern matching often substitutes for algorithms

Application-level consequences

- temporal reasoning errors
- math word problem failures
- difficulty with unsolvable or faulty MWPs

Elementary numerical failures propagate into many downstream reasoning tasks.

Embodied reasoning: from text to perception to action

1D: text-based physical commonsense & physics and scientific reasoning

2D: anomaly detection & spatial reasoning & visual physical reasoning

3D: affordance/planning & spatial/tool-use reasoning & safety/autonomy

Language competence does not imply grounded physical understanding.

Cross-cutting root causes across the map

- next-token prediction favors pattern completion
- transformer structure introduces asymmetries and order sensitivity
- passive text learning lacks grounding and feedback
- internal world models remain weak
- memory, planning, and composition remain fragile

Many seemingly different failures arise from a smaller set of recurring causes.

How are we doing in mitigating them?

Synthesis and future directions

What is needed, i.e. missing from the current landscape?

- deeper root-cause analysis
- unified persistent failure benchmarks
- failure injection into general benchmarks
- dynamic / private / event-driven evaluation
- more interactive and multi-turn settings

The field needs a mature science of reasoning failure analysis.

The taxonomy becomes a fork in the road

INFORMAL REASONING

Open-ended behavior
Context-dependent norms
No universal correctness checker

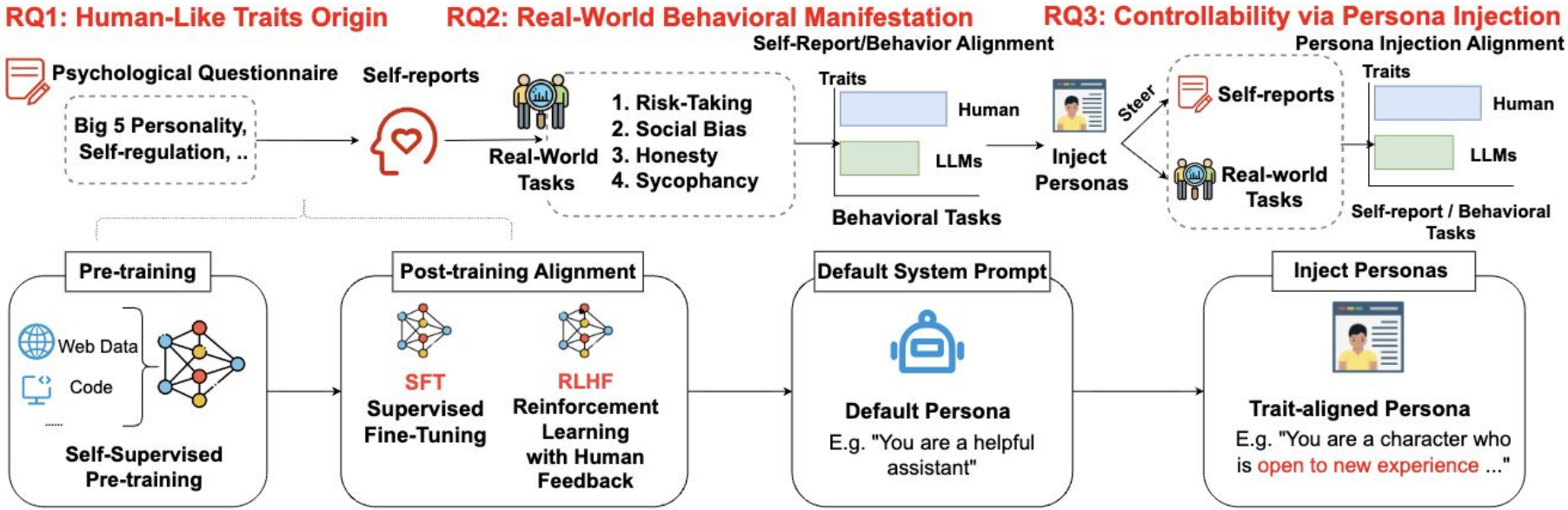
⇒ richer behavioral and interactive evaluation

FORMAL REASONING

Explicit rules and semantics
Machine-checkable artifacts
Reliable verification signals

⇒ verification inside the reasoning loop

Informal: toward behavioral & interactive evaluation



Informal: toward behavioral & interactive evaluation

Response-centered Evaluation

Final Response
The primary evaluated artifact is final answer, label, generated text or predicted actions.

Output Evaluator
Usually score a final answer against a reference, rubric or judge.

Output Level Judge
Score / label assigned to response such as correctness, similarity, pass/fail ...

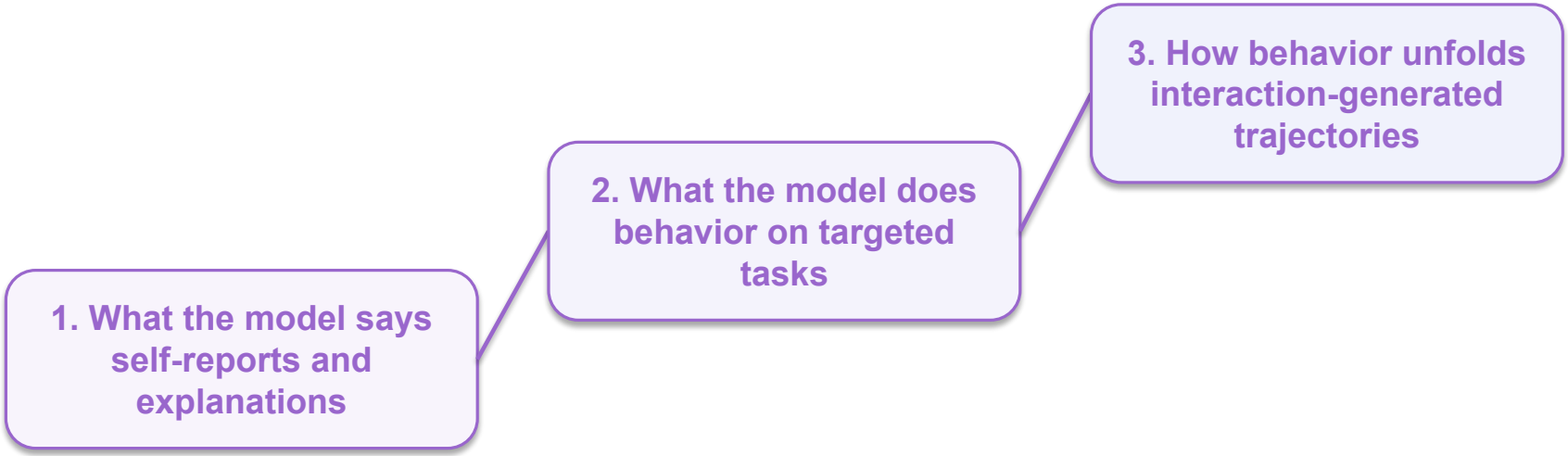
Interactive Evaluation

Interaction Trajectory
The primary evaluated artifact is trajectory such as observations, tool calls or state transitions.

Trajectory Evaluator
Map trajectory to judgments over interaction quality, efficiency, safety or recovery...

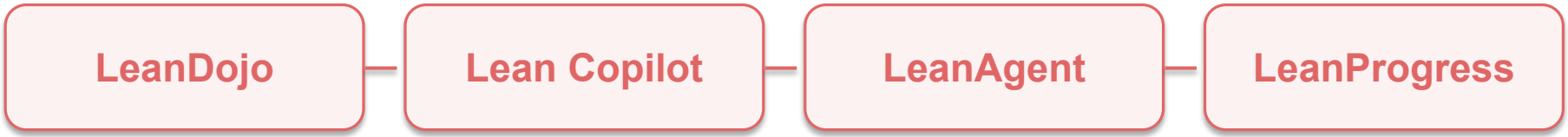
System Level Judge
The evaluation reflects outcome, process, recovery, safety, and efficiency beyond final response quality

An evidence ladder for informal reasoning



The strength of an evaluation claim must match the richness of its evidence.

Formal: the verifier closes the loop



Open the environment

framework · models · data · benchmark

Bring models into the workflow

tactic suggestions · premises · proof search

Adapt as knowledge grows

continual learning adaptation

Give search a direction

proof-progress prediction long-context ability

Next: bringing verifiability deeper into reasoning systems

Zooming out: reasoning is a feedback problem

Informal
construct meaningful behavioral evidence

Formal
exploit exact machine-checkable verification

Embodied
learn from grounded, consequential interaction

Reasoning as creative as human intuition and as dependable as formal logic

Stanford University Caltech Carnegie Mellon University 上海人工智能实验室 Shanghai Artificial Intelligence Laboratory

From Reasoning Failures to Reliable Intelligence

Understanding, Evaluating, and Building Better LLM Reasoning Systems

Future models should not only reason better — they should also fail better.

Questions?