LLM Reasoning and Agents

CS 288: Advanced Natural Language Processing

“Reasoning” Models

- Going all-in on serialized inference

- CoT is a powerful approach for getting better performance on any instruction-tuned model

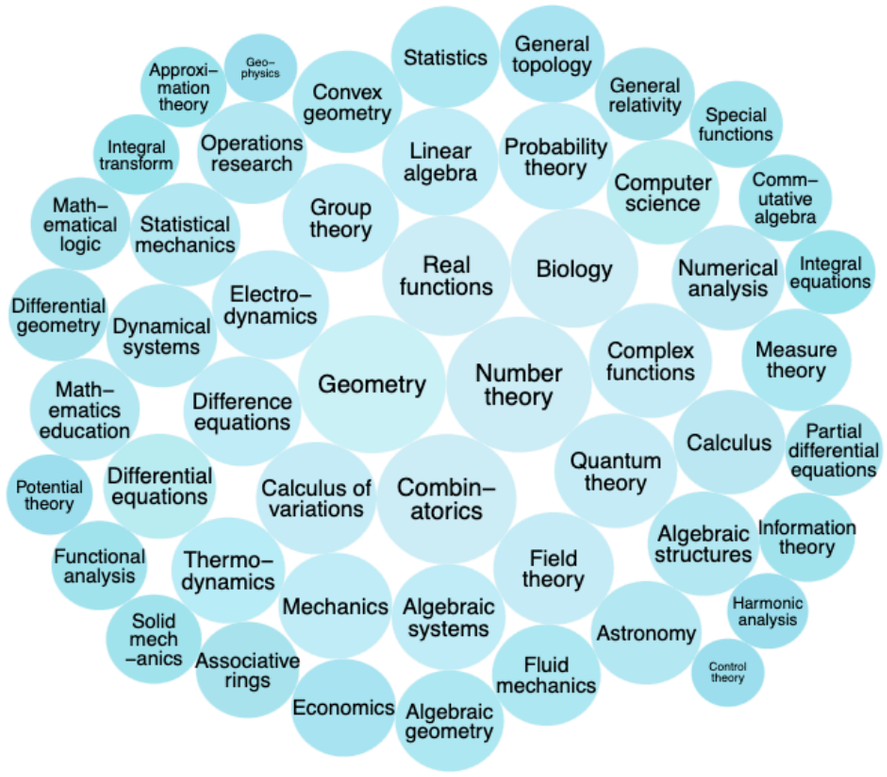
- A “reasoning” model is explicitly trained to generate CoT

Self-Taught Reasoner (STaR)

- During training, iteratively:
  - Map from questions to answers with CoT using existing policy, resulting in <question, CoT, answer> tuples
  - Modify tuples based on correctness:
    - If the answer is correct, keep it as is
    - If the answer is incorrect, prompt policy to map from question and correct answer to a new CoT (“rationalization”)
  - Fine-tune policy on modified tuples
- Why not train only using rationalizations?

Supervised Fine-Tuning

- Main idea: curate a dataset of high-quality reasoning traces, then fine-tune models on these traces
- s1: 1,000 reasoning traces sampled from existing models
- Inference-time compute budget control



How many r in raspberry? **Question**

Let's break down the process of counting the letter 'r' in the word "raspberry" ...
* First letter: 'r' - This is an 'r', count = 1.
* Second letter: 'a' - Not an 'r', count remains 1 ...
* Sixth letter: 'e' - Not an 'r', count remains 1.
* Seventh letter: 'r' - This is an 'r', count = 2.
* Eighth letter: 'y' - Not an 'r', count remains 2 ...
The number of 'r's in "raspberry" is 2.
Wait, let's re-read the question carefully. It asks "How many r in raspberry?" ... * r - a - s - p - b - e - r - r - y ... * First 'r' ... * Second 'r' ... * Third 'r' ... Count = 3 ... **Reasoning trace**

My initial answer of 2 was incorrect due to a quick reading of the word. **Final Answer:** The final answer is $\boxed{3}$ **Response**

Muennighoff et al. 2025

Error Propagation in Sequential Decision Making

- Core challenge in dynamic systems: executing a policy may result in states for which it has poor estimates of what's best to do, e.g.:
  - Stuck in a corner
  - Repeated the same action over and over
  - Opened an app it was never trained on that looks similar to the correct app
- Problem: even SFT'd models haven't been trained on what to do, because experts rarely get into these situations

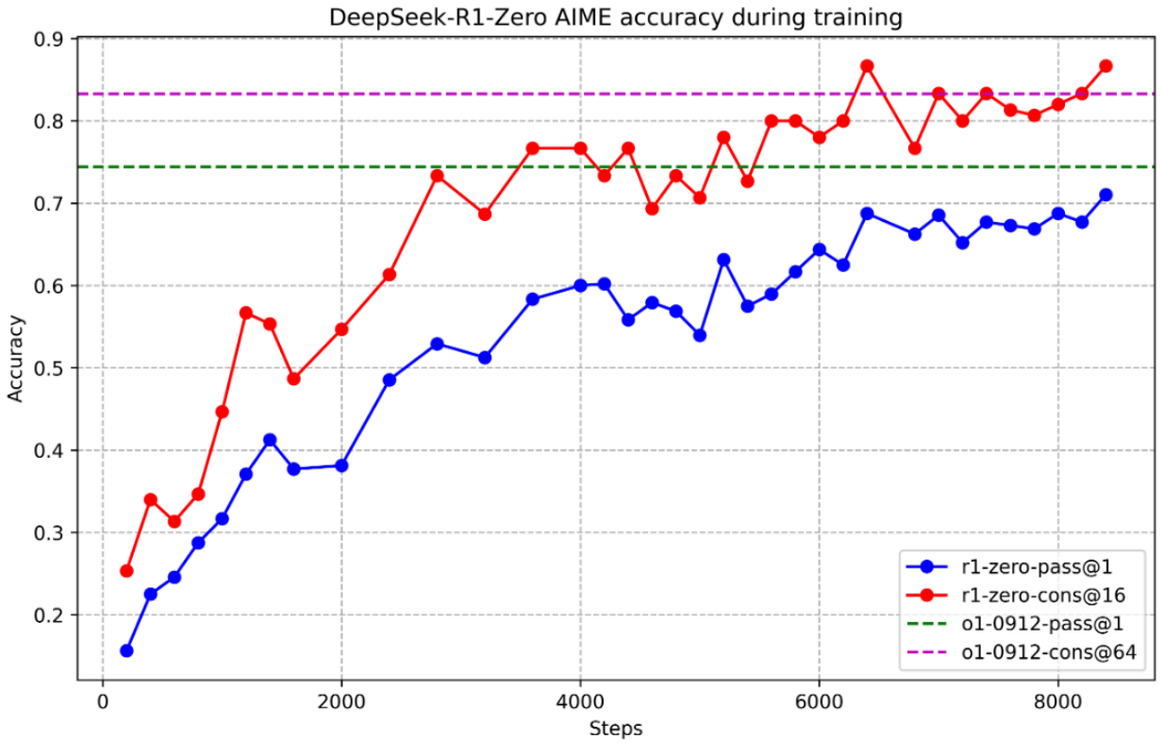
Learning from Exploration

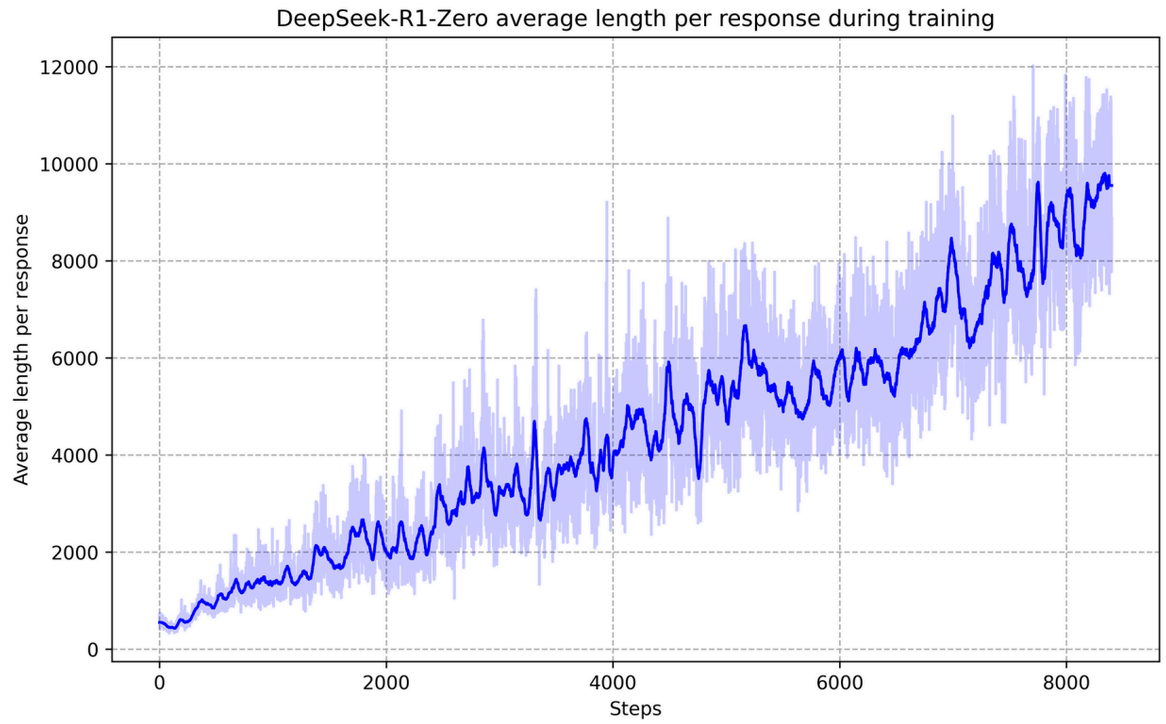
- Instead of learning only from high-quality demonstrations, where local decisions are ~optimal, let's learn from what states the policy is likely to encounter at inference time

- **Really high-level overview:** learn by
  - Sampling from the current policy
  - Evaluating the quality of these samples
  - Optimizing the policy towards or away from the sample, depending on its quality

DeepSeek-R1

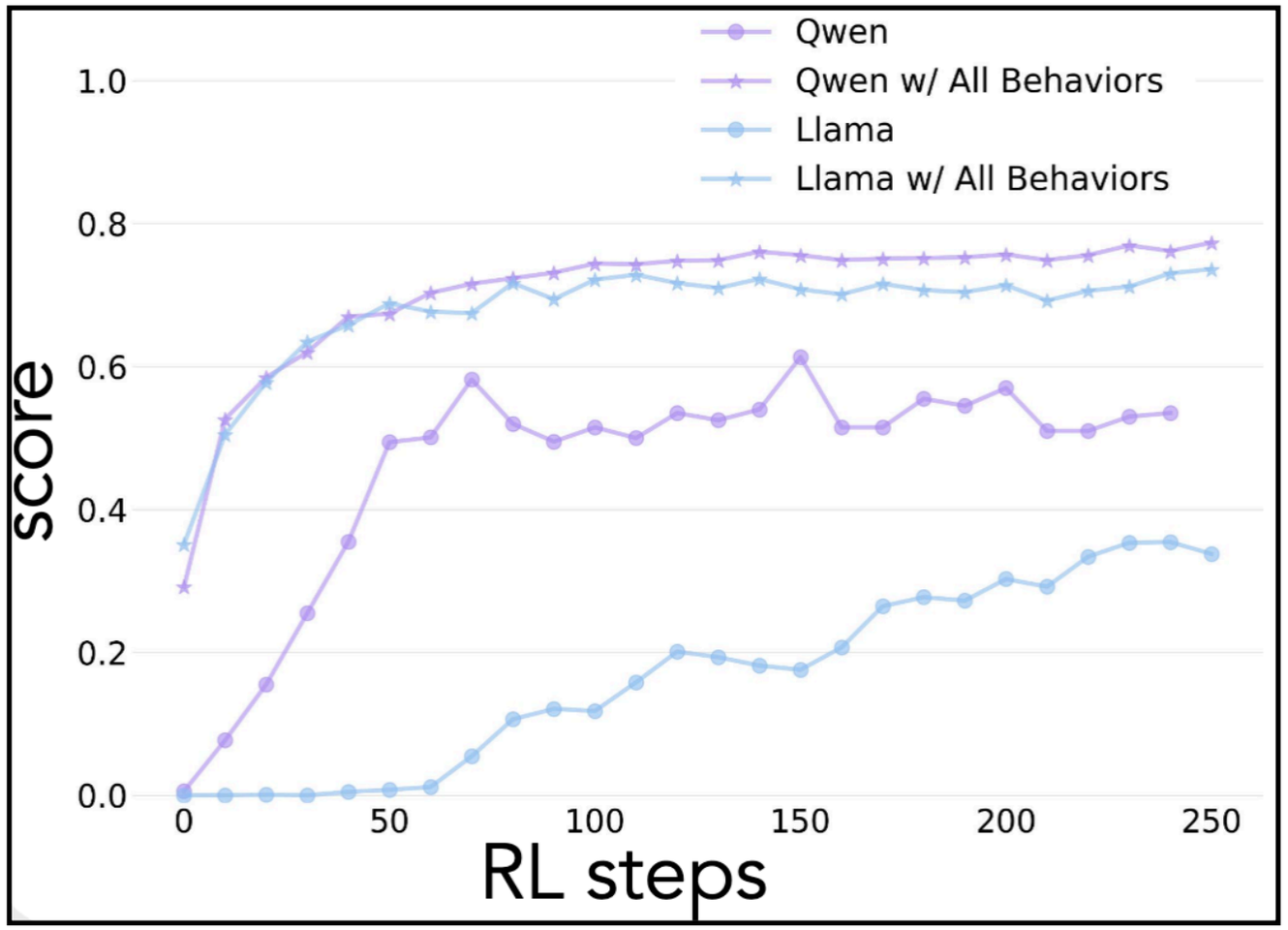
- During training, iteratively:
  - Map from questions to answers with CoT using existing policy, resulting in <question, CoT, answer> tuples
  - Assign a score to each tuple based on correctness
  - Optimize policy towards tuples with high score





Four Critical Reasoning Behaviors

- Self-verification
  - “Let me check my answer”
- Subgoal setting
  - “Let’s try to get a multiple of 10”
- Backtracking
  - “Let’s try a different approach, what if we…”
- Backward chaining
  - “Working backwards, 24 is 8 times 3”



But: faithfulness and interpretability problems remain

Some Remaining Questions in “Reasoning” Models

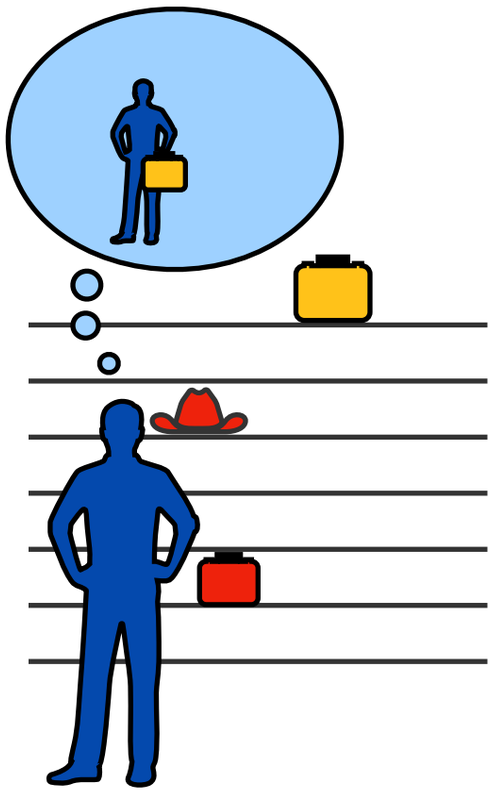
- How to train most effectively — online RL? pretraining data like s1?
- Under what conditions do different “reasoning strategies” “emerge”?
- How faithful are learned long CoT to actual reasoning processes? How interpretable are the CoT traces?
- How well does learning on one task generalize to other tasks?
- How can we more effectively take advantage of inference-time compute?

Slides adapted from Graham Neubig

Agents

- What is an agent?
- Definition we will take here: an entity that takes actions that influence the state of the **dynamic** system in which it is embodied
  - Simple vision+language tasks (captioning, entailment): non-agentic, because context is a static image
  - What kinds of contexts are dynamic?

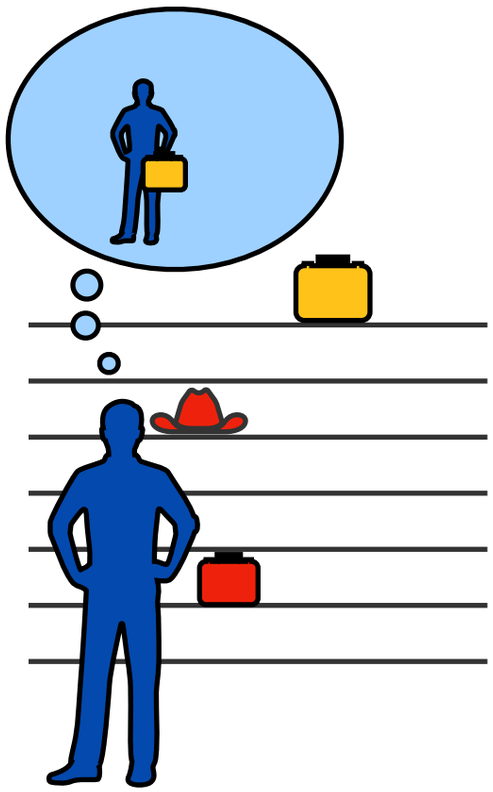
Partially Observable Markov Decision Processes



<u>My goal:</u>
Get the suitcase

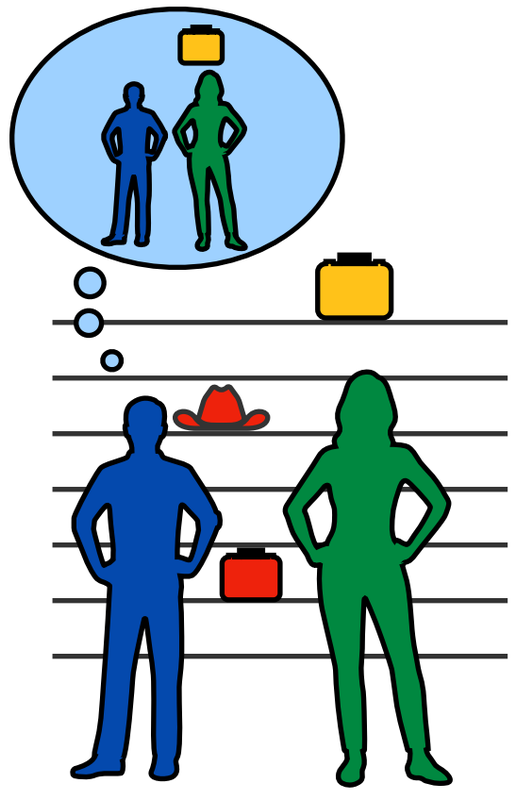
but… I'm too short

Partially Observable Markov Decision Processes



My goal:
Get the suitcase

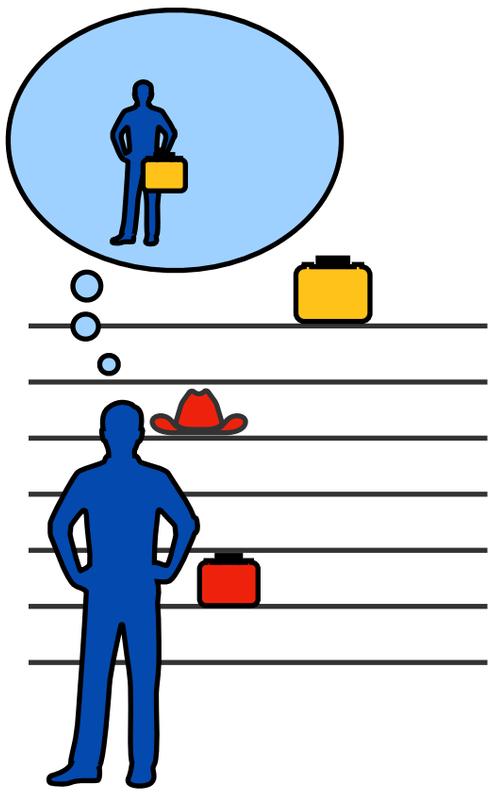
but… I'm too short



Observation:
Green person can reach it

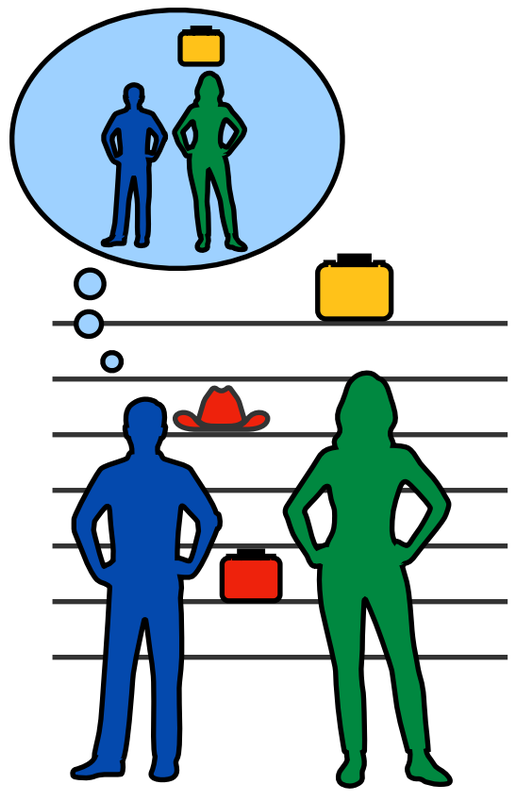
Optimal action:
Green person gives it to me

Partially Observable Markov Decision Processes



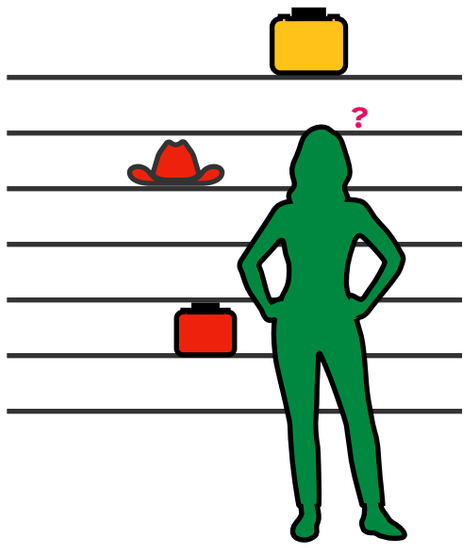
My goal:
Get the suitcase

but… I'm too short



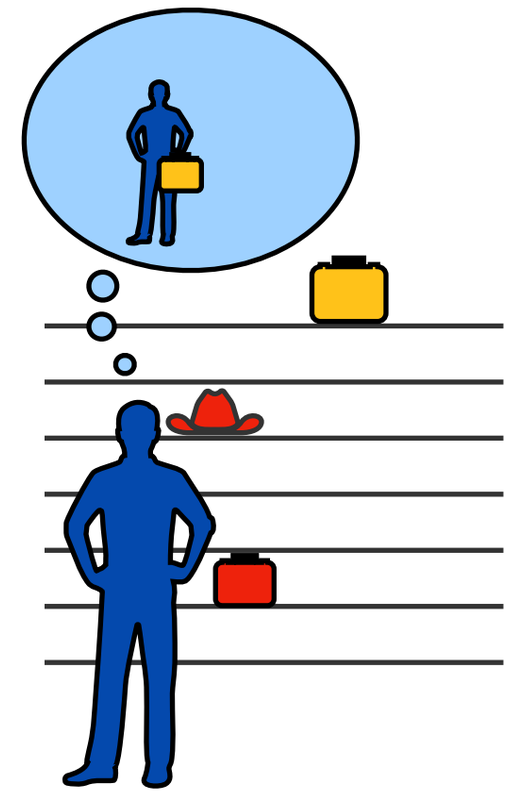
Observation:
Green person can reach it

Optimal action:
Green person gives it to me



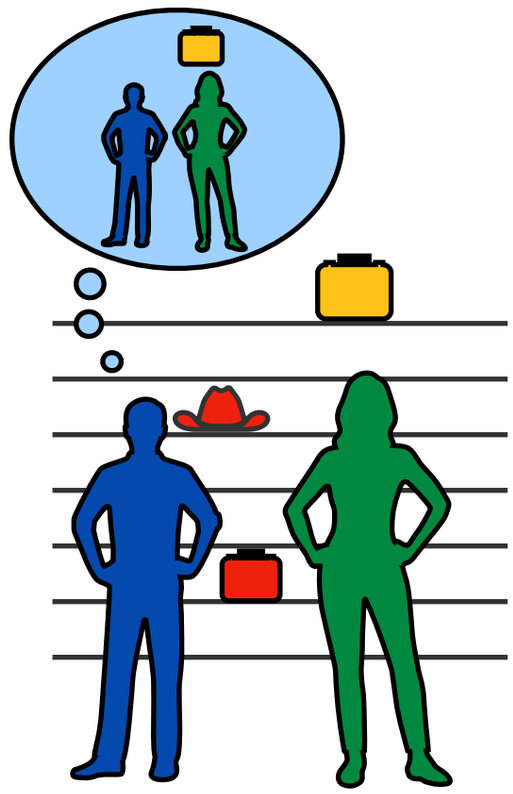
Belief:
Green person doesn't know my goal

Partially Observable Markov Decision Processes



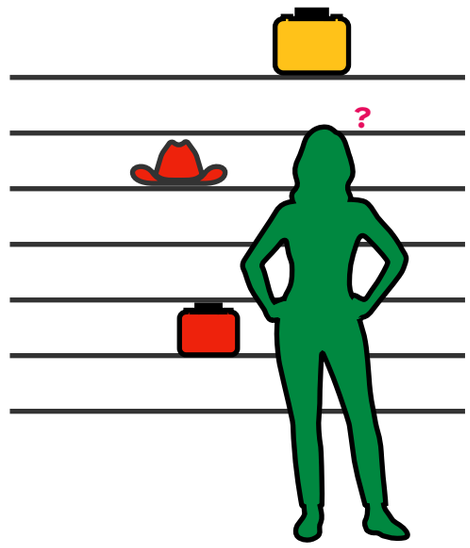
My goal:
Get the suitcase

but… I'm too short

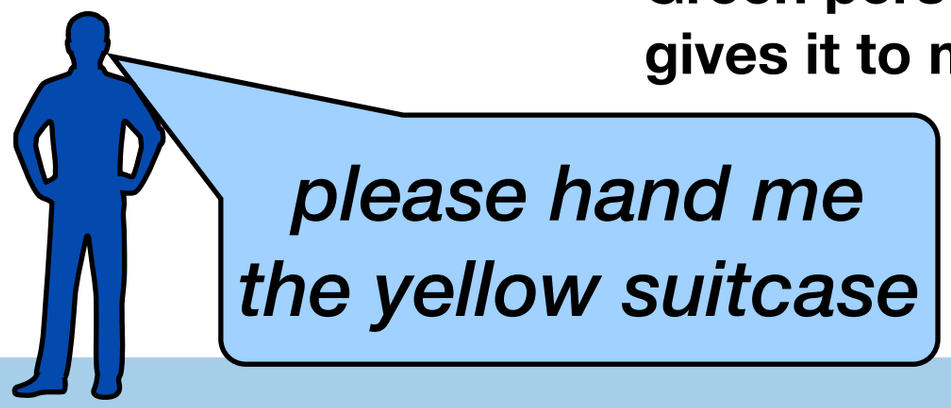


Observation:
Green person can reach it

Optimal action:
Green person gives it to me



Belief:
Green person doesn't know my goal



Task:
Language Generation

States:

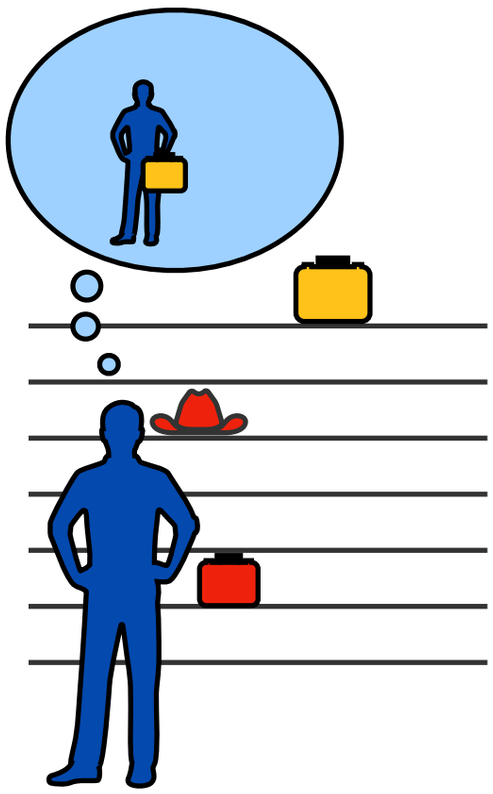
Observations:

Actions:

Transition function:

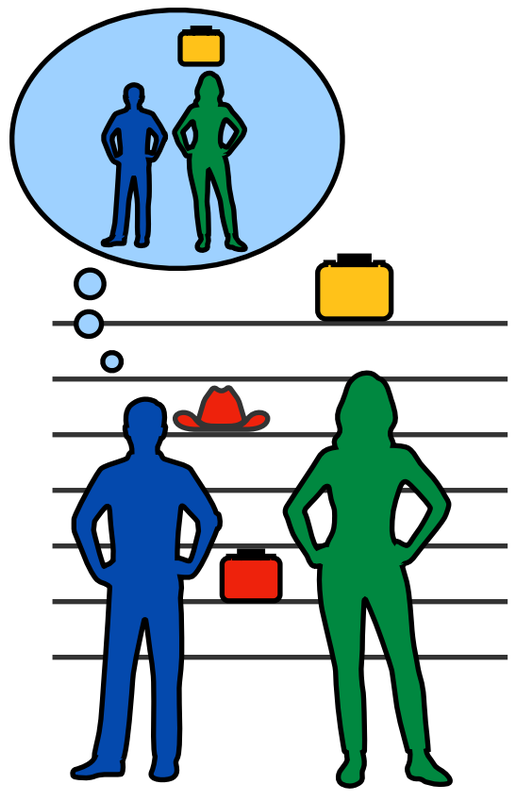
Reward function:

Partially Observable Markov Decision Processes



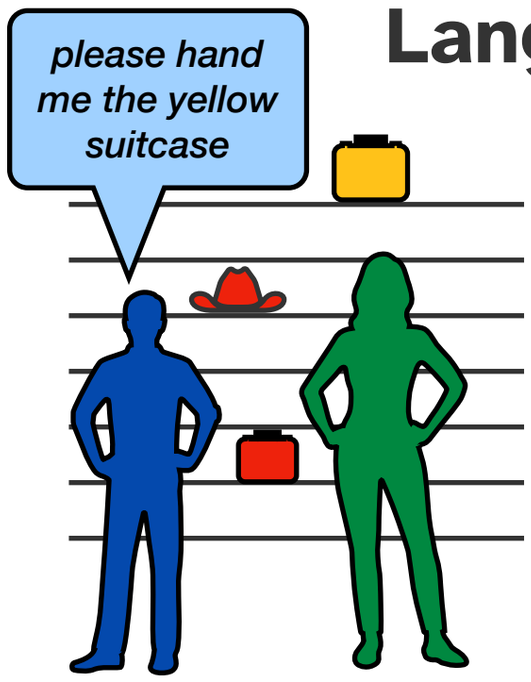
My goal:
Get the suitcase

but… I'm too short



Observation:
Green person can reach it

Optimal action:
Green person gives it to me



Belief:
Green person doesn't know my goal

Task:
Language Understanding

States:

Observations:

Actions:

Transition function:

Reward function:

POMDP:
Grounded Instruction Following

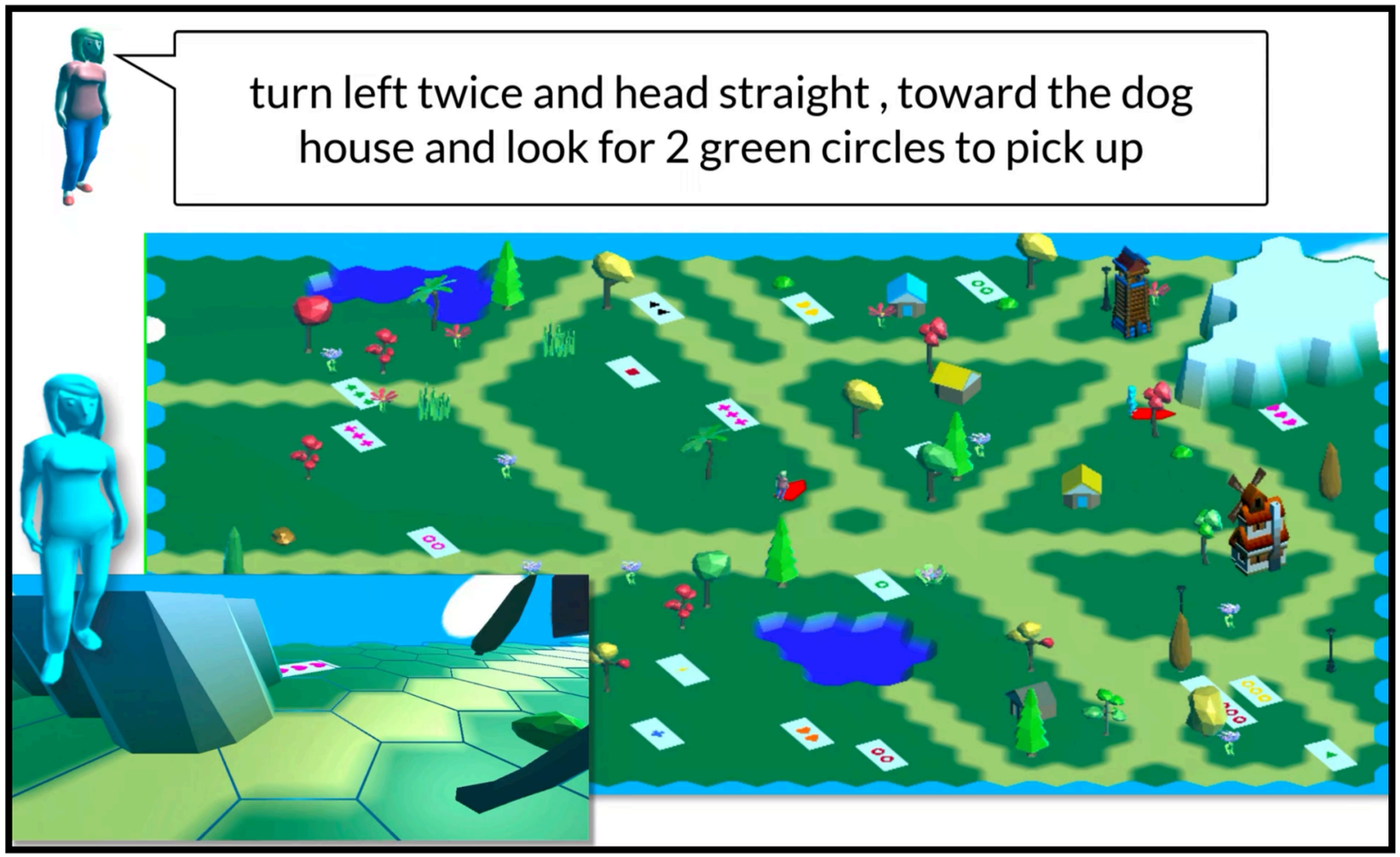
States:

Observations:

Actions:

Transition function:

Reward function:



CerealBar, Suhr et al. 2019

POMDP:
Software Engineering

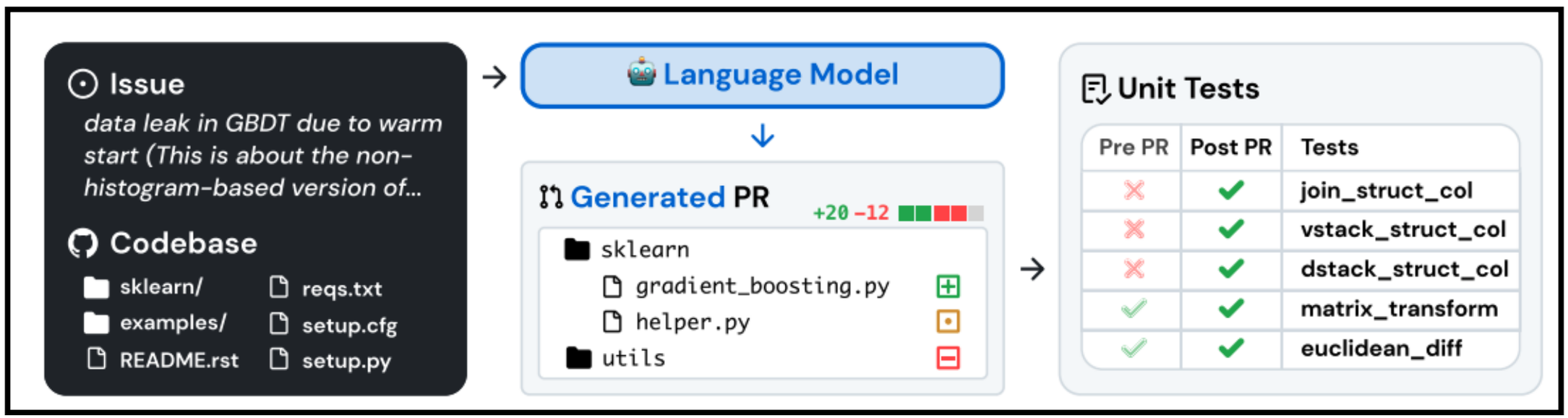
States:

Transition function:

Observations:

Reward function:

Actions:



SWE-Bench, Jimenez et al. 2023

POMDP:
Device Control

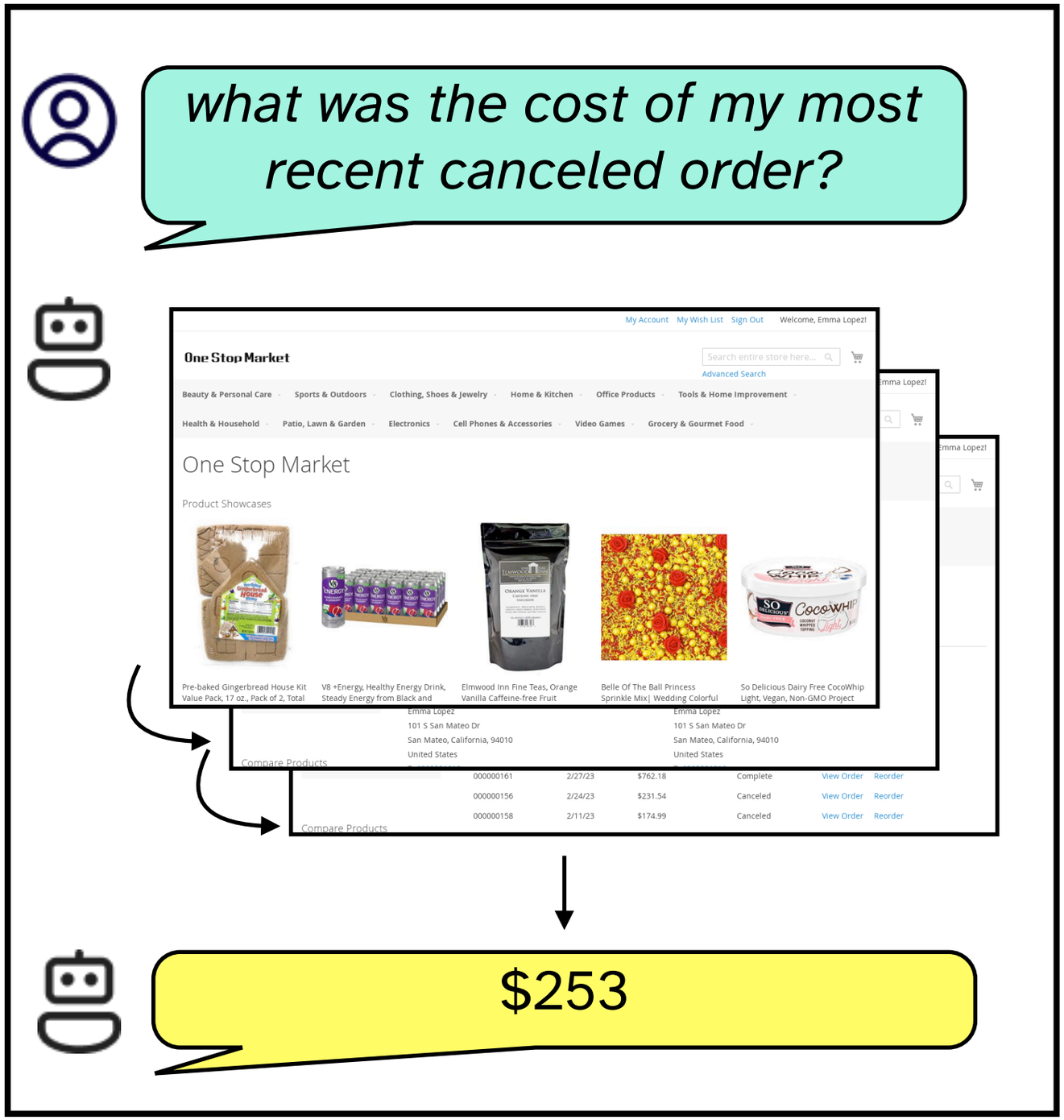
States:

Observations:

Actions:

Transition function:

Reward function:



WebArena, Zhou et al. 2023

Agent Challenges

- Input and output space is domain-specific, and pretrained LMs may struggle to generalize to new domains
- Policy must reason about environment dynamics
  - Sequential decision-making: chance for error propagation!
  - How have we addressed error propagation in the past? (hint: remember RLHF?)

Domain Generalization via Tools

- Instead of expecting model to do everything itself, **build tools that we can prompt it to use**

- Benefits:
  - Tools can more easily provide guarantees of success (e.g., calculator)
  - When prompted right, instruction-tuned LLMs can reliably call tools and use their results
  - LLMs don’t need to be fine-tuned to successfully take actions in target domains

Slides adapted from Zora Wang

Structured Prompting and "Tool-Calling"

- Why are we expecting the models to do arithmetic directly? Why not just give them a calculator?
- Main idea: prompt LMs to "call" tools, e.g., by interleaving language output with calls to a calculator:

```
A: The bakers started with 200 loaves.
loaves_baked = 200

They sold 93 in the morning and 39 in the afternoon.
loaves_sold_morning = 93
loaves_sold_afternoon = 39

The grocery store returned 6 loaves.
loaves_returned = 9

The answer is
answer = loaves_baked - loaves_sold_morning -
         loaves_sold_afternoon + loaves_returned
```

prompt, model text output, model program output

Gao et al. 2022

Kinds of Tools

- **Perception:** collect new data from the environment
- **Action:** exert actions by changing the environment state
- **Computation:** offload computation onto an actual computer

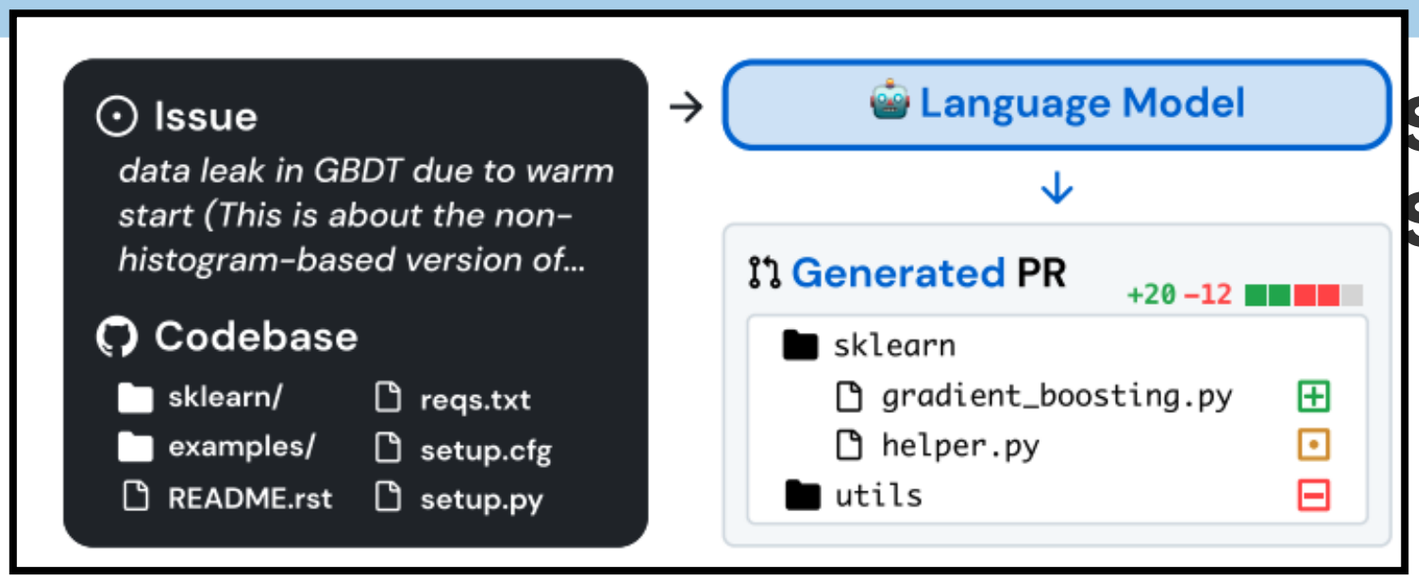
Kinds of Tools

Edit this image so there are twice as many cats on the porch.

Scenario: Embodied Instruction Following

- **Perception:** collect new data from the environment
- **Action:** exert actions by changing the environment state
- **Computation:** offload computation onto an actual computer

Kinds of Tools

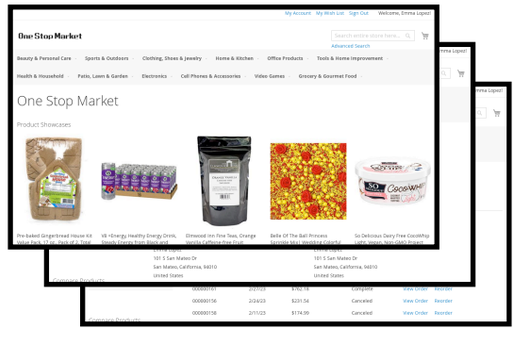


Scenario: Automated Software Engineering

- **Perception:** collect new data from the environment
- **Action:** exert actions by changing the environment state
- **Computation:** offload computation onto an actual computer

Kinds of Tools

what was the cost of my most recent canceled order?



Scenario: Device Control

- **Perception:** collect new data from the environment
- **Action:** exert actions by changing the environment state
- **Computation:** offload computation onto an actual computer

Prompt-Based Agents

- Defining the POMDP ourselves allows us to deploy LLMs as agents in a variety of domains
  - Observation space
  - Action space and transition function (tools)
- The right prompt design and inference method can further improve agent performance, e.g.:
  - Prompt the model to plan its actions, and revise its plan, before executing the plan
  - Sample multiple possible trajectories, then choose the one with the highest potential for success with some reward model
- Supervised fine-tuning with expert demonstrations can also help!

Error Propagation in Sequential Decision Making

- Core challenge in dynamic systems: executing a policy may result in states for which it has poor estimates of what's best to do, e.g.:
  - Stuck in a corner
  - Repeated the same action over and over
  - Opened an app it was never trained on that looks similar to the correct app
- Problem: even SFT'd models haven't been trained on what to do, because experts rarely get into these situations

Reinforcement Learning for Language Agents

- During learning:
  - Sample a trajectory from the current policy
  - Assign the trajectory a reward
  - Optimize the policy to maximize the reward

- What we need:
  - States — including a diverse set of instructions!
  - Actions and transition function — an executable environment!
  - Reward function — some kind of automatic reward function!

Case Study: Software Engineering

States: current state of the code

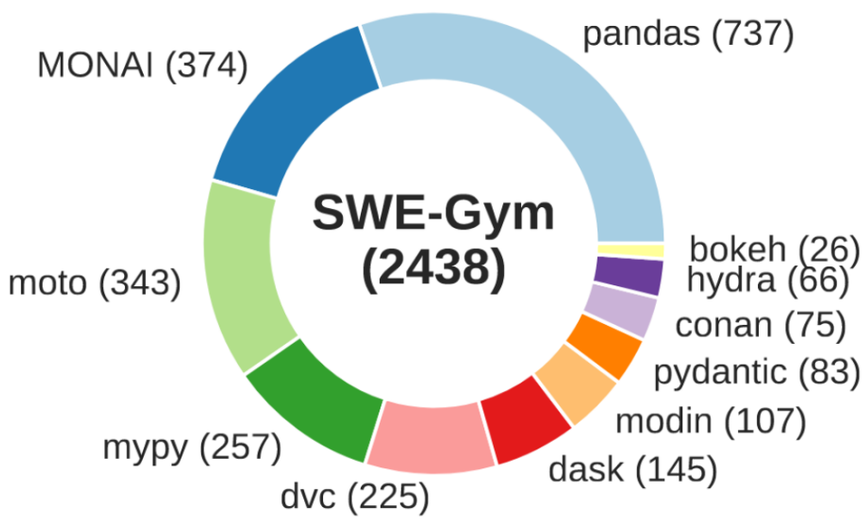
Observations: what's in the shell/editor

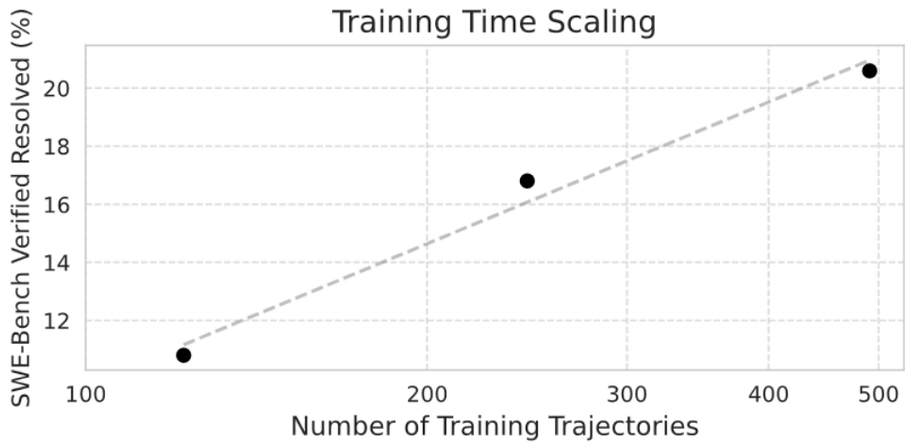
Actions: edit the code or execute a command

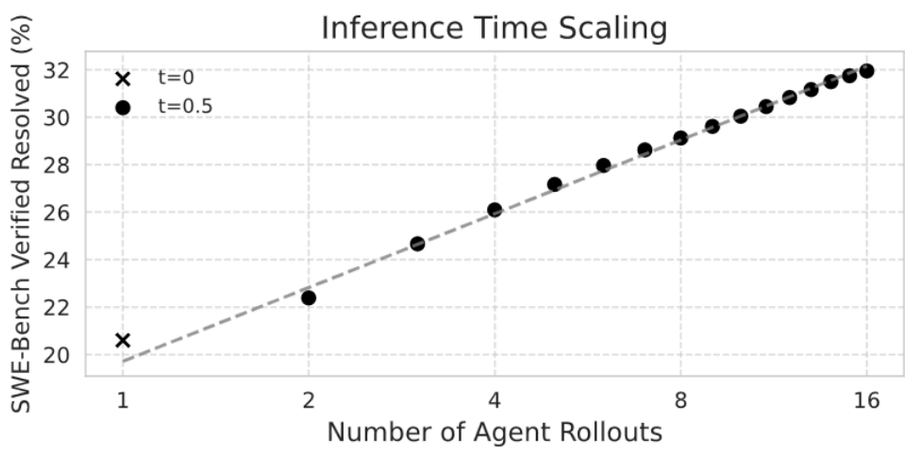
Transition function: apply the edits or execute command

Reward function: whether the current code passes unit tests

SWE-Gym: public training environments for software engineering

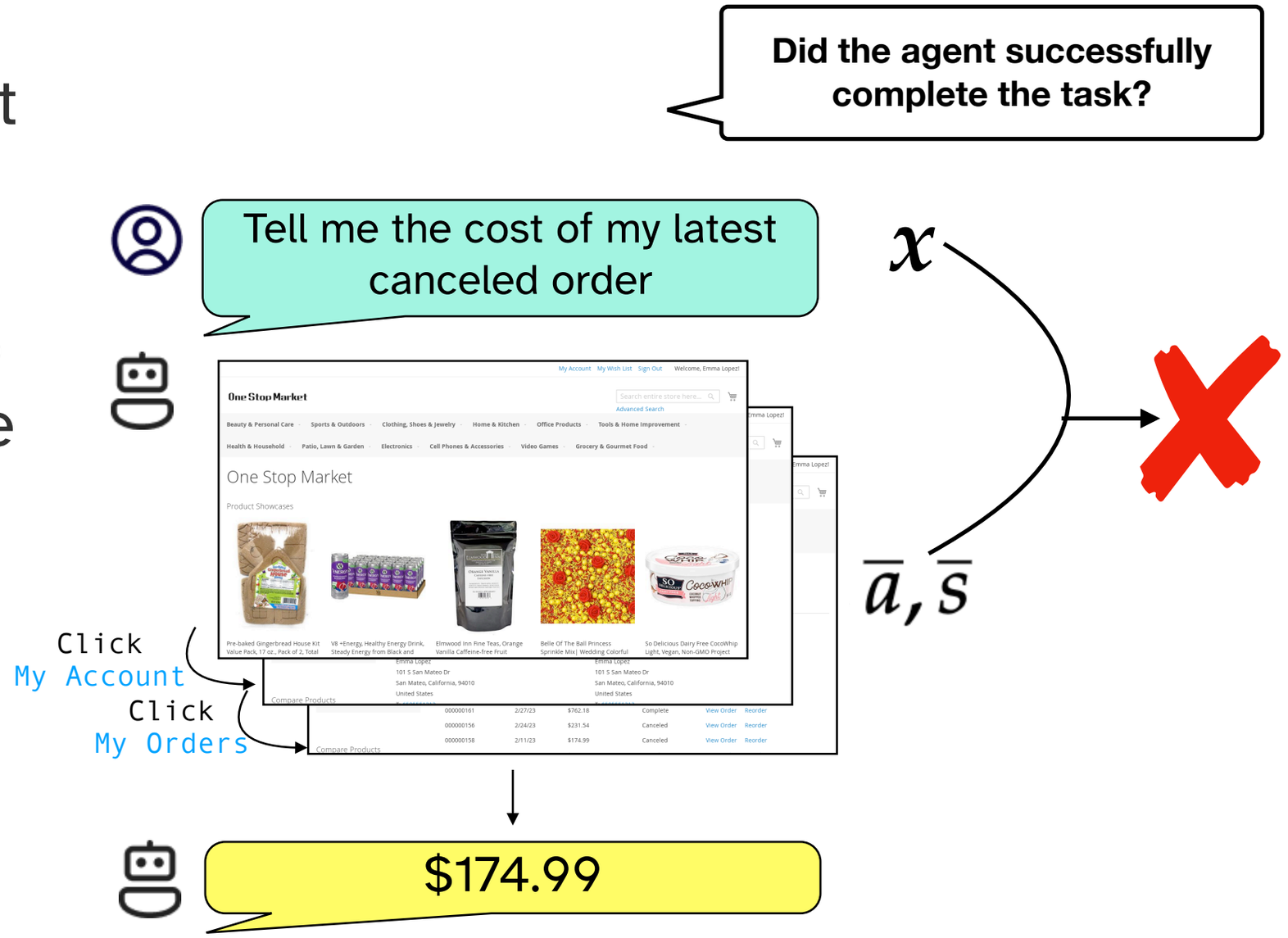






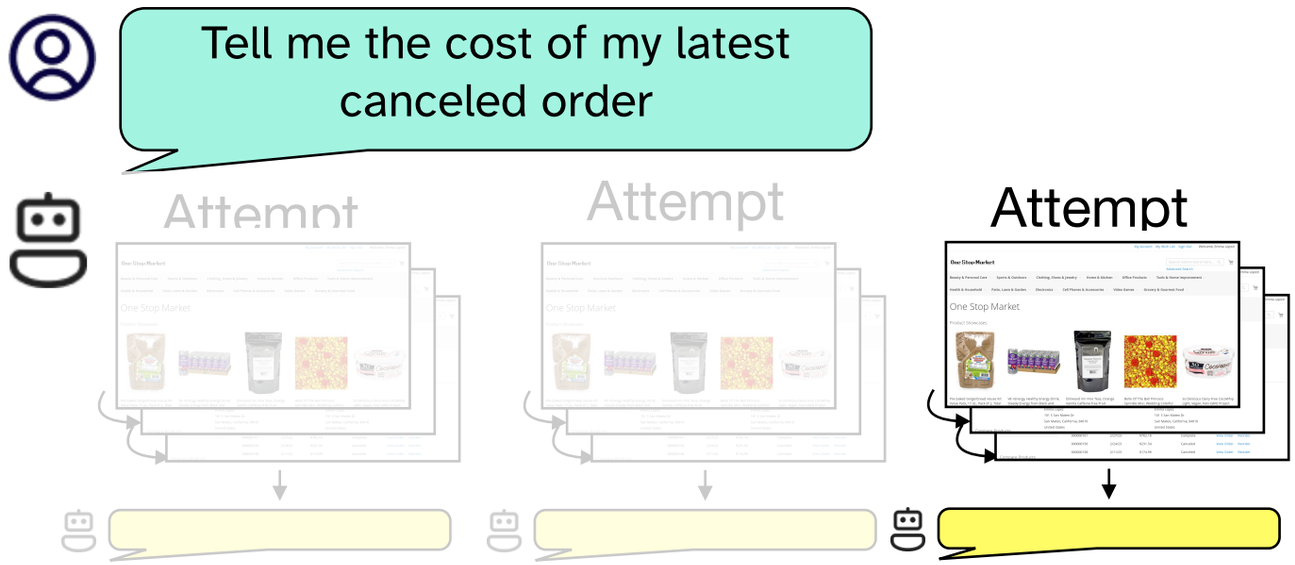
Automating evaluation

- Writing high-quality test cases is expensive
- This limits our ability to learn from more diverse sets of tasks
- Instead, how can we leverage automatic evaluators during training?



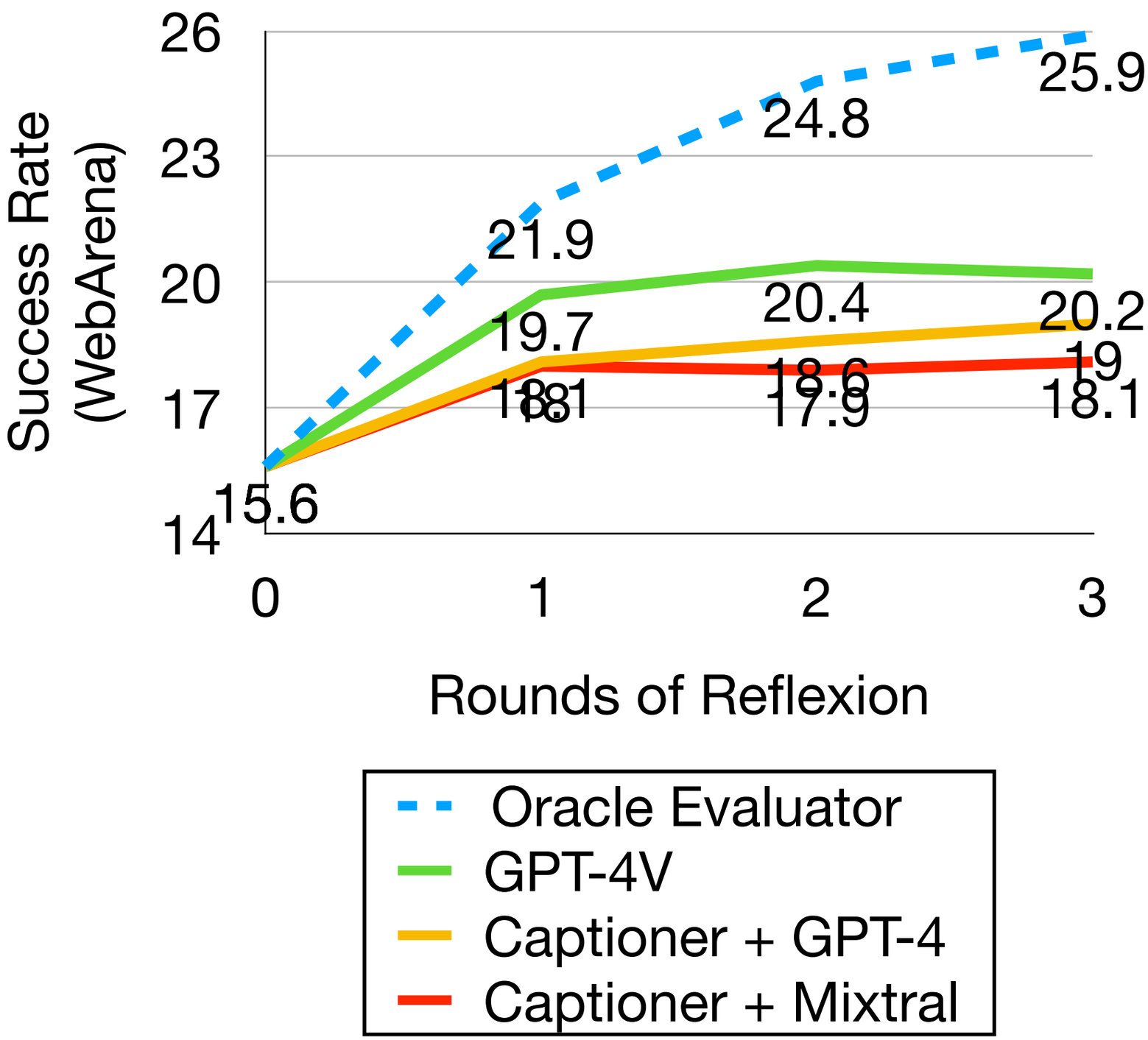
Improving embodied agents with automated evaluators

At inference time with Reflexion (Shinn et al. 2023)



Automated evaluator:

Improving embodied agents with automated evaluators



At inference time with Reflexion (Shinn et al. 2023)

Up to 29% relative improvement over SOTA using automatic evaluator

(16% w/ open weight models)

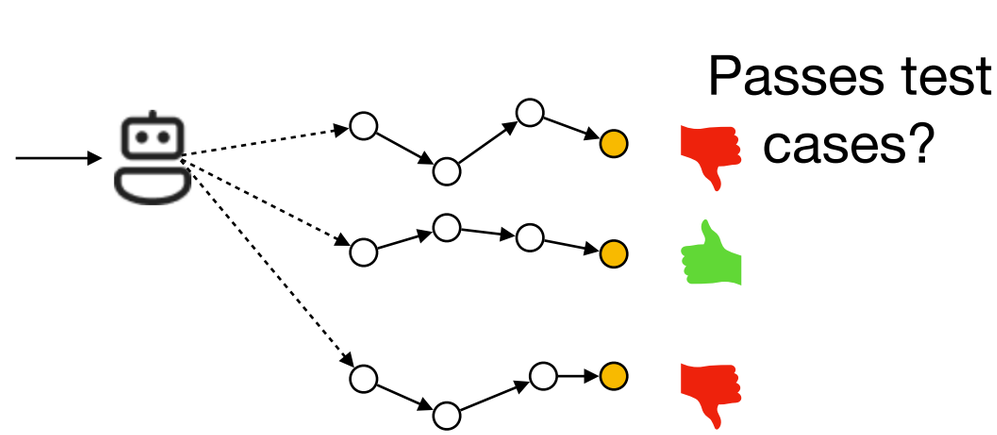
Prompt-based automatic evaluators

Improving embodied agents with automated evaluators

At inference time with reranking

Training time

issue description (training set)

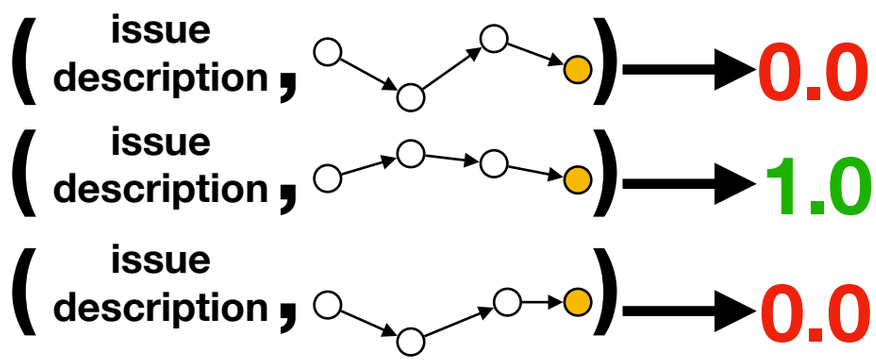


Sample *n* patches

Improving embodied agents with automated evaluators

At inference time with reranking

Training time

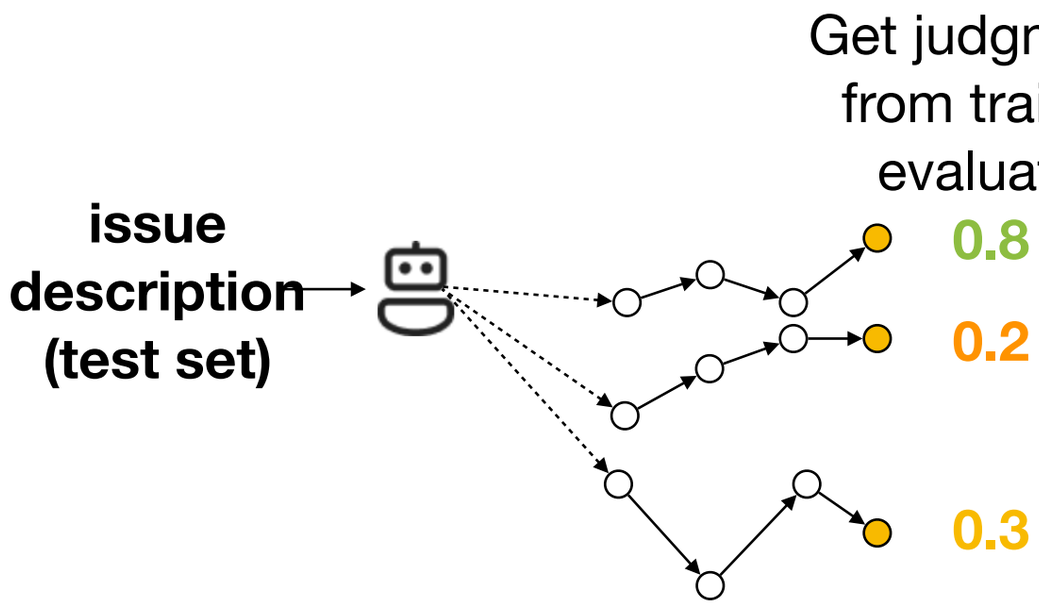


Training examples for automated evaluator

Improving embodied agents with automated evaluators

At inference time with reranking

Inference time

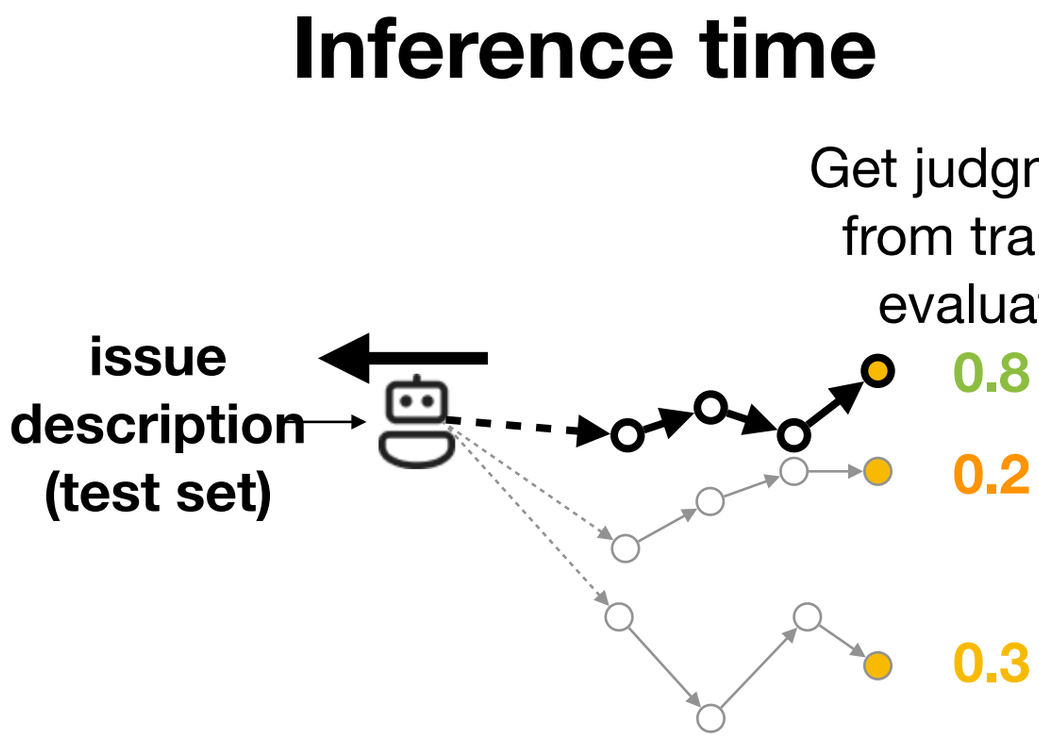


Sample *n* trajectories

Pan et al. 2025, ICML

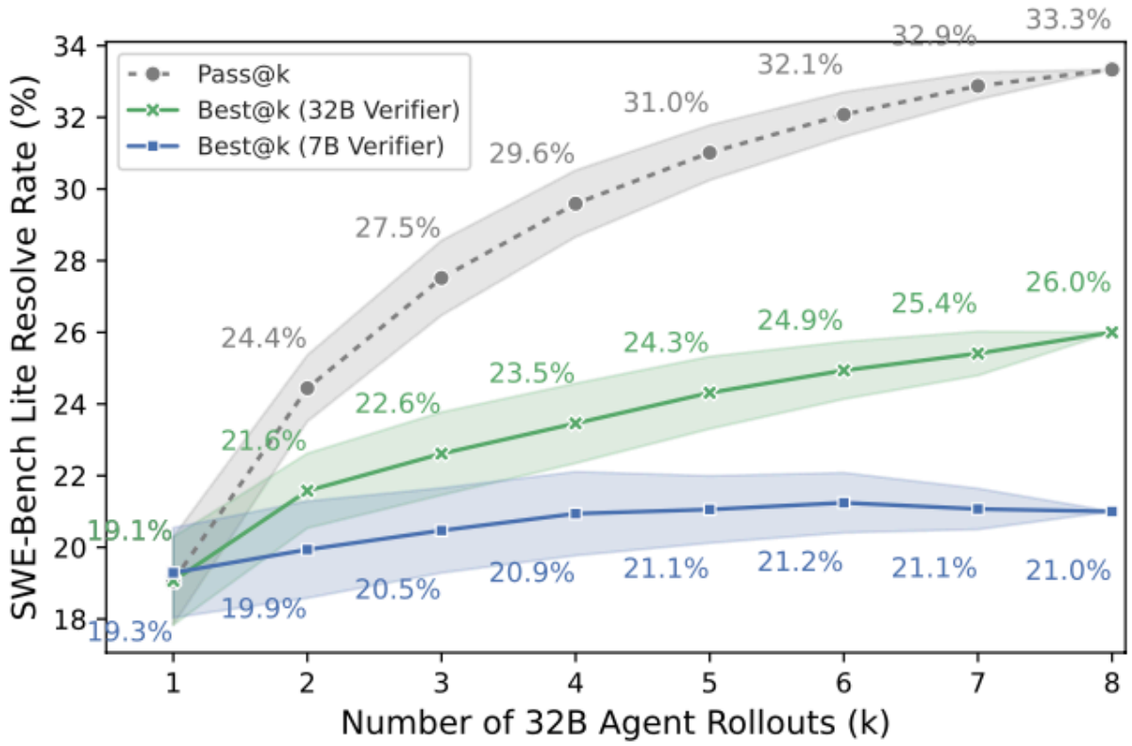
Improving embodied agents with automated evaluators

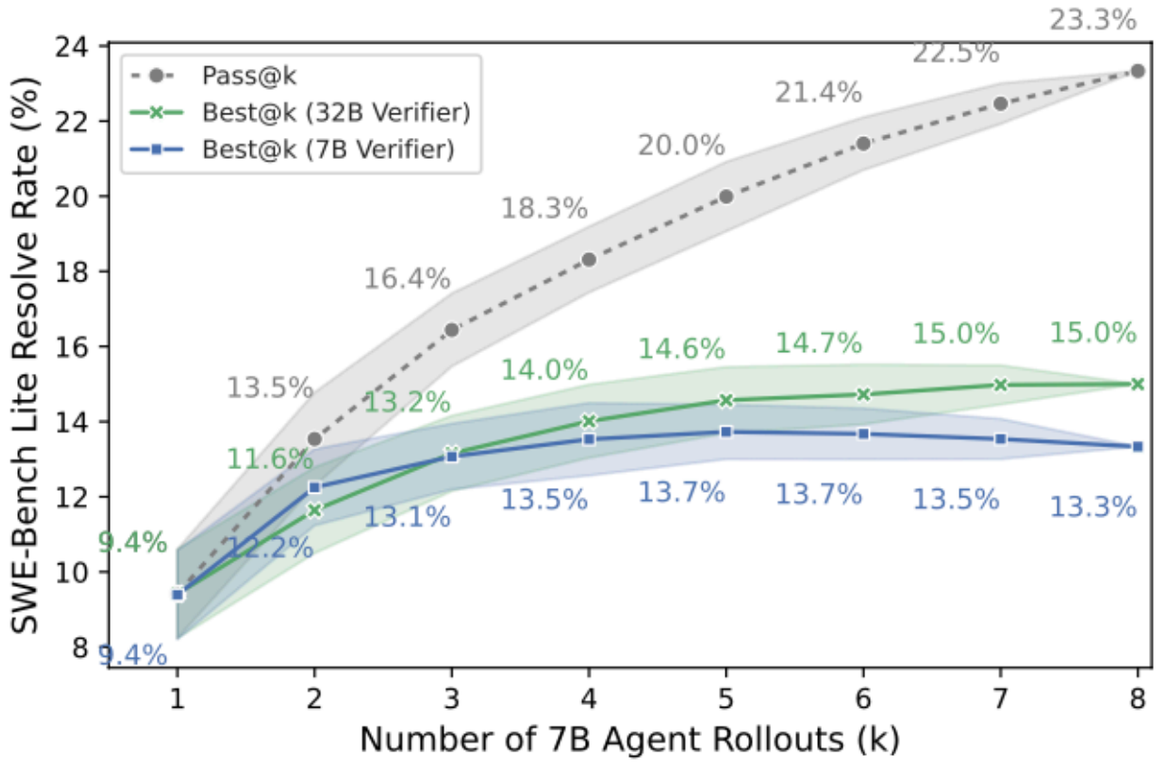
At inference time with reranking



Improving embodied agents with automated evaluators

At inference time with reranking



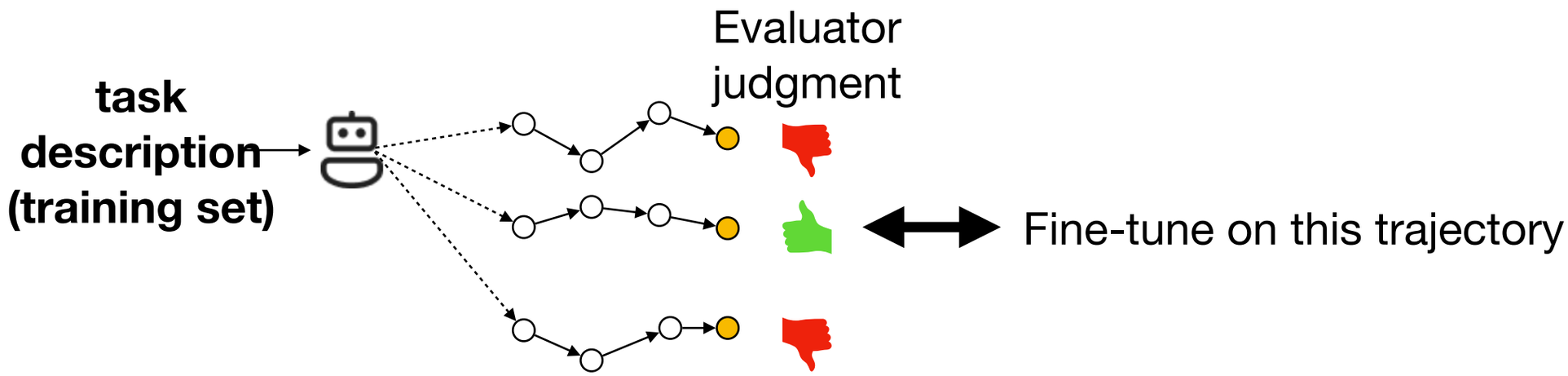


Pan et al. 2025, ICML

Improving embodied agents with automated evaluators

At training time
as a reward function

Filtered behavior cloning



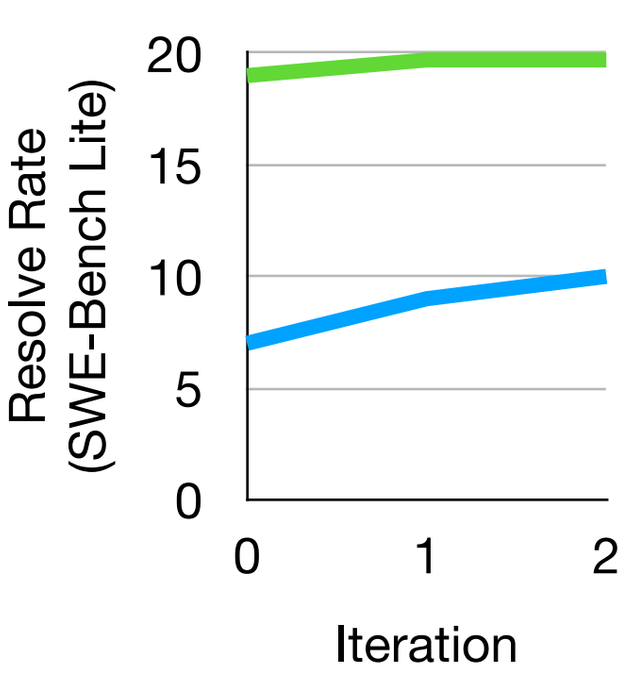
Sample *n* trajectories

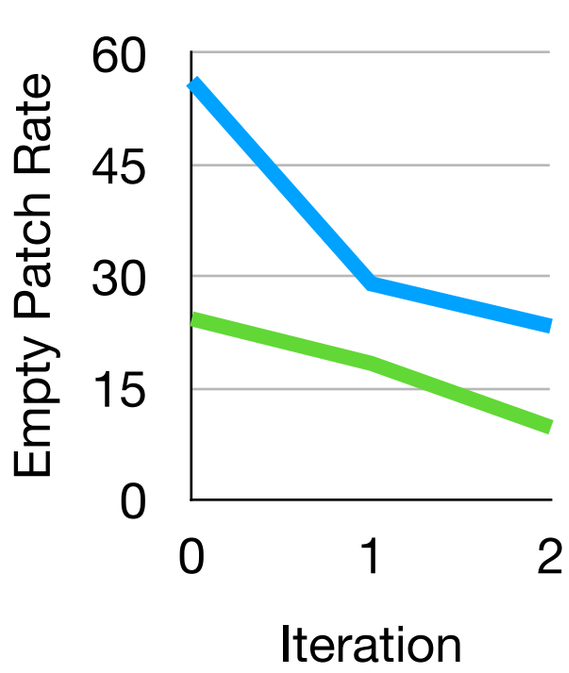
Improving embodied agents with automated evaluators

At training time
as a reward function

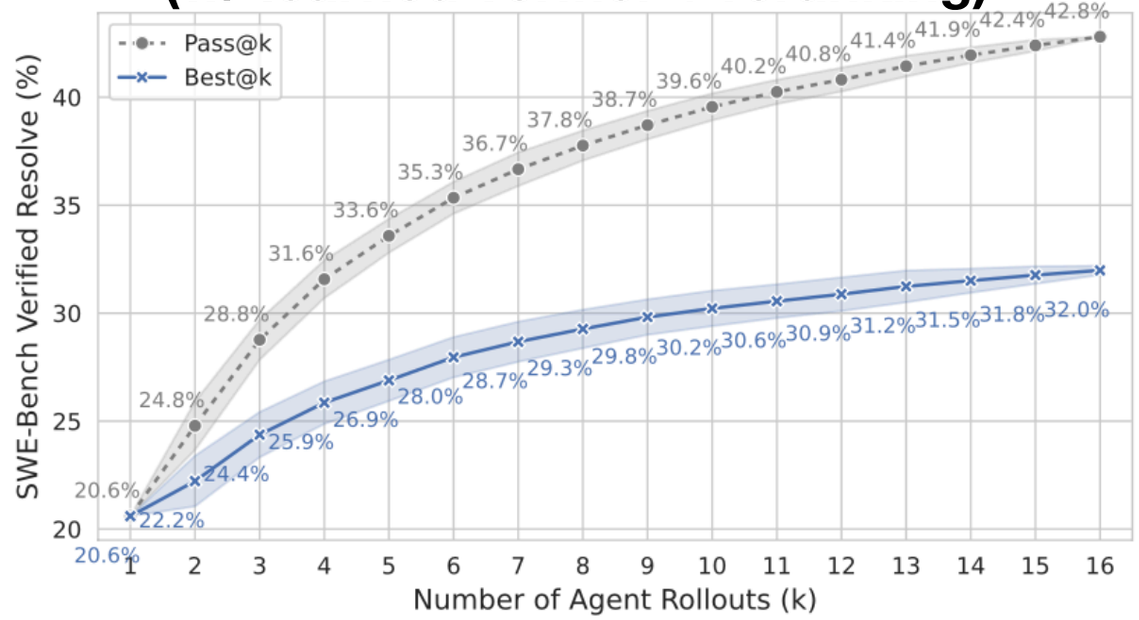
With **SWE-Gym** using filtered behavior cloning

- 7B Model
- 32B Model





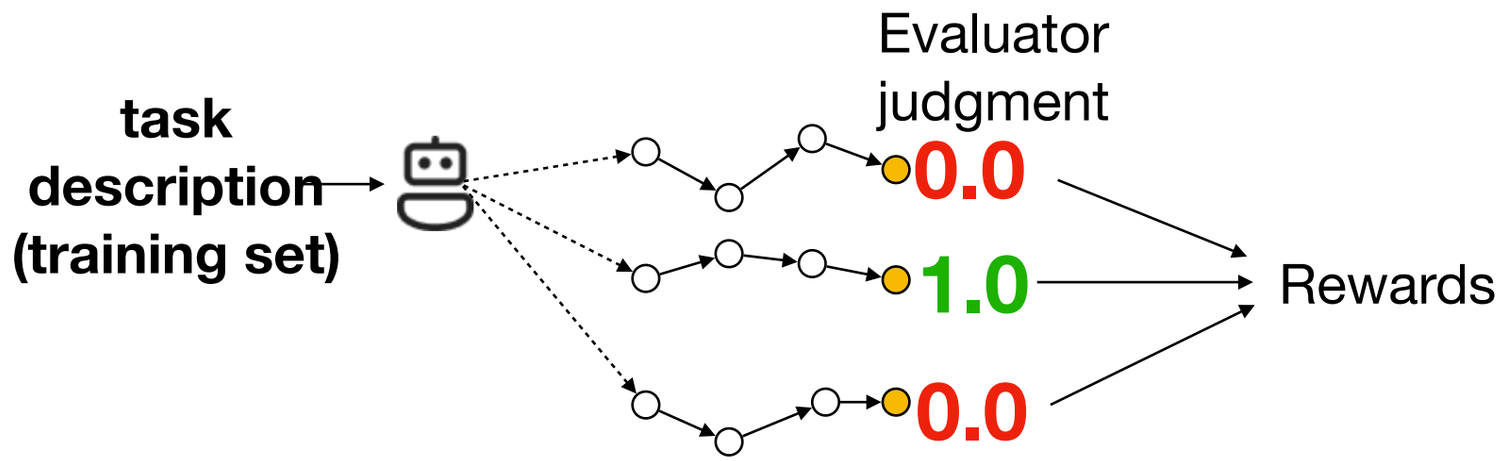
(w/ learned verifier + reranking)



Improving embodied agents with automated evaluators

At training time
as a reward function

Reinforcement learning



Sample *n* trajectories

Improving embodied agents with automated evaluators

At training time
as a reward function

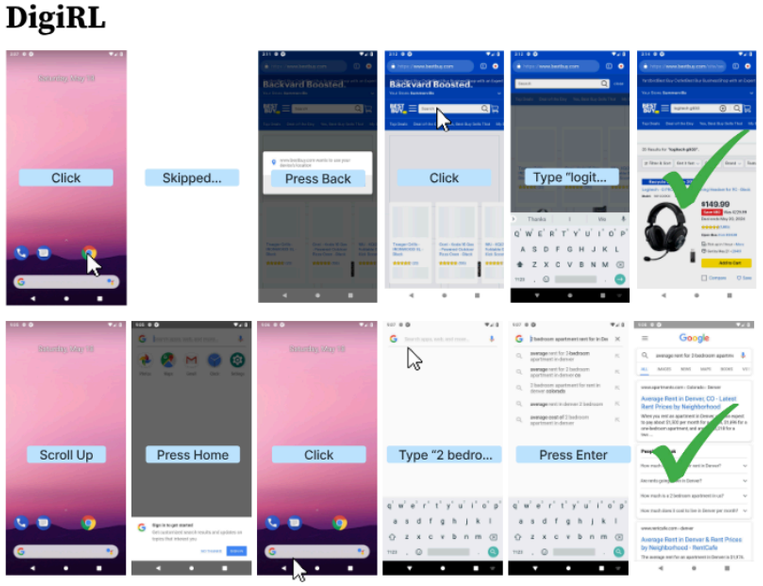
DigiRL

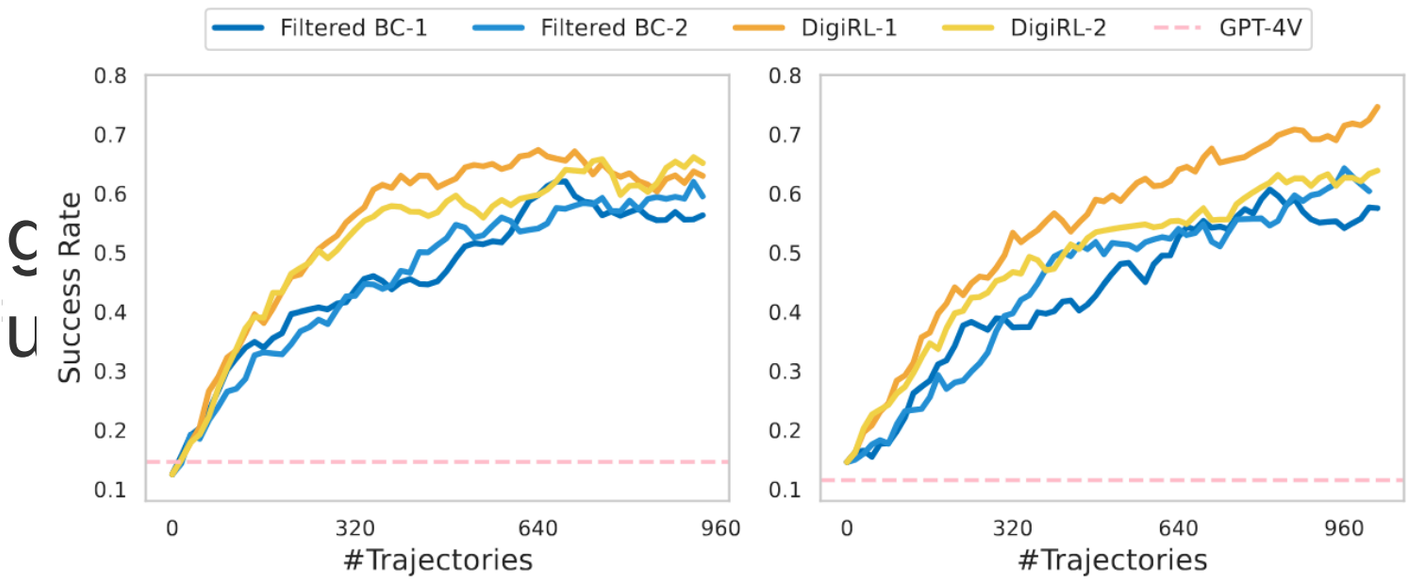
WebShop

Go to bestbuy.com, search for “logitech g933”

General

How much does a 2 bedroom apartment rent for in Denver?

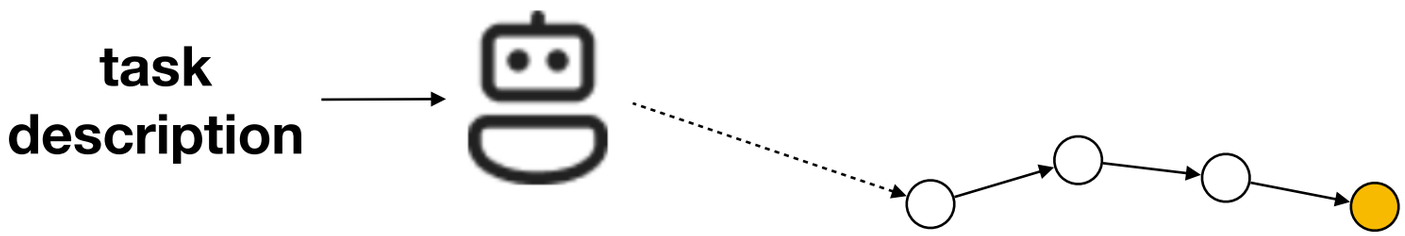




More efficient reasoning through task decomposition

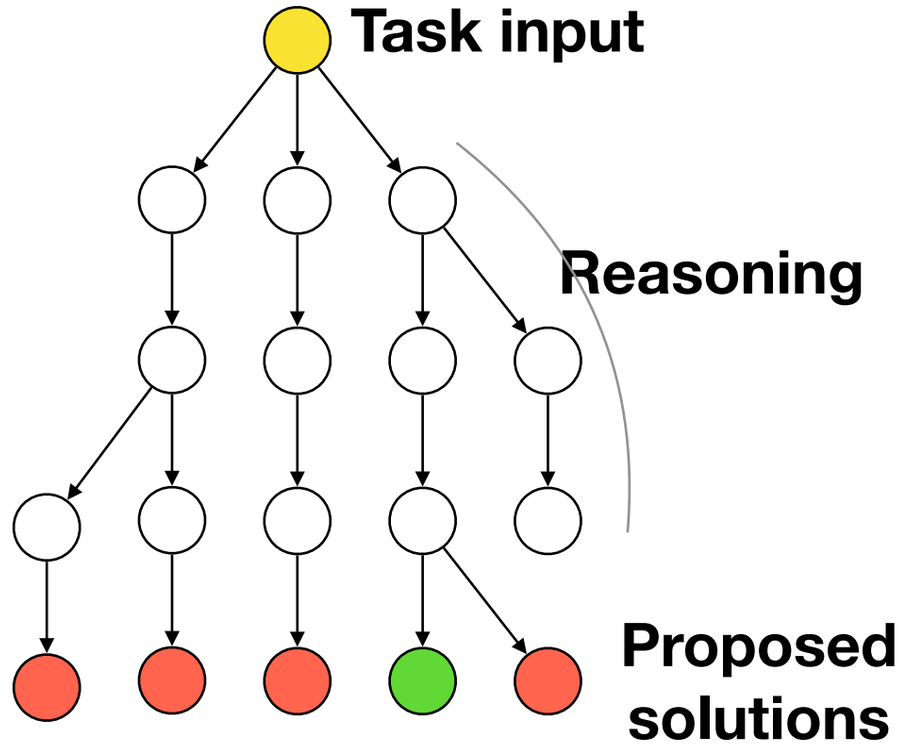
Are our tasks really sequential?

And if they aren't, are there design assumptions we don't have to make?



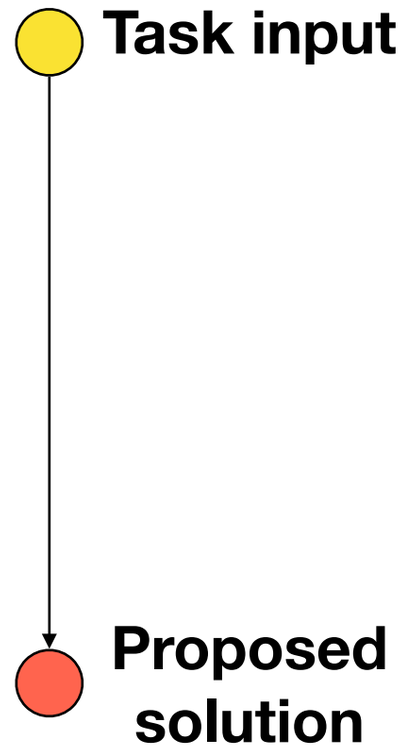
Non-sequential reasoning

- Software engineering
  - Investigate multiple possible causes of a bug
  - Refactor all files according to a new standard
- Math reasoning
  - Different strategies for approaching a problem
  - Working backwards from multiple possible solutions
- General natural language reasoning
  - Do a literature search for different potential explanations of a phenomenon
  - Hierarchical structure of composing a piece of text



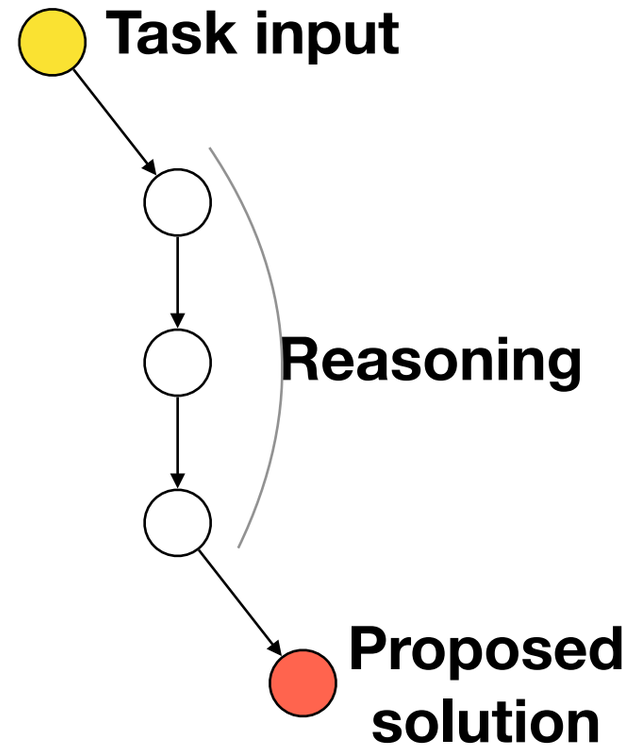
Types of reasoning models

- Inference structures
  - Zero-shot



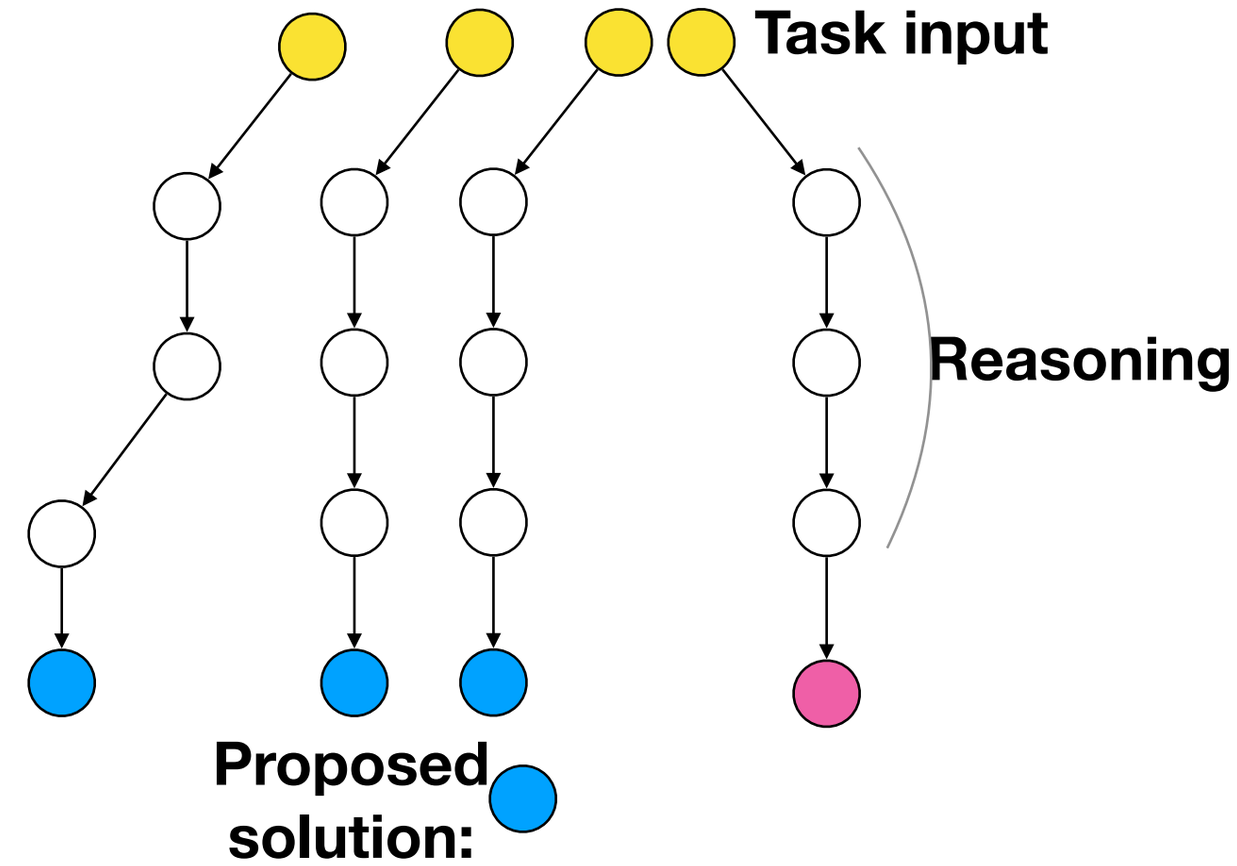
Types of reasoning models

- Inference structures
  - Zero-shot
  - Chain-of-thought



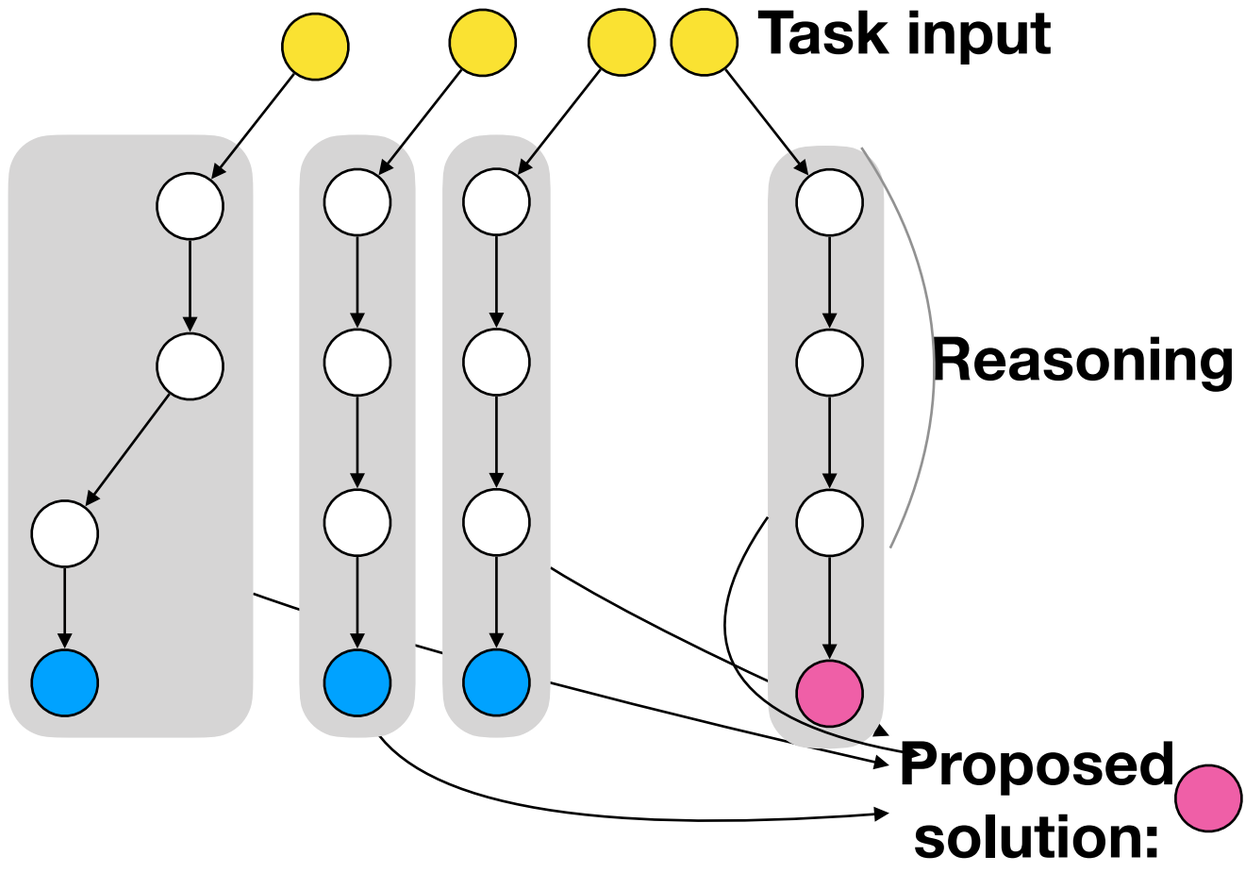
Types of reasoning models

- Inference structures
  - Zero-shot
  - Chain-of-thought
  - Parallel approaches (e.g., consensus)



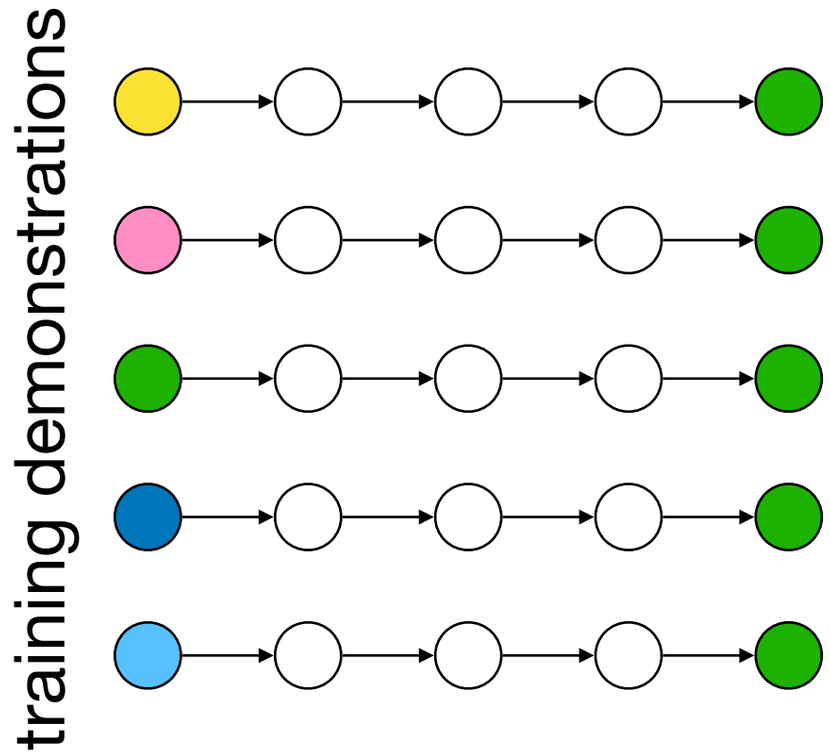
Types of reasoning models

- Inference structures
  - Zero-shot
  - Chain-of-thought
  - Parallel approaches (e.g., consensus)
  - Structured inference (e.g., tree-of-thoughts, debate, inference with verifiers)



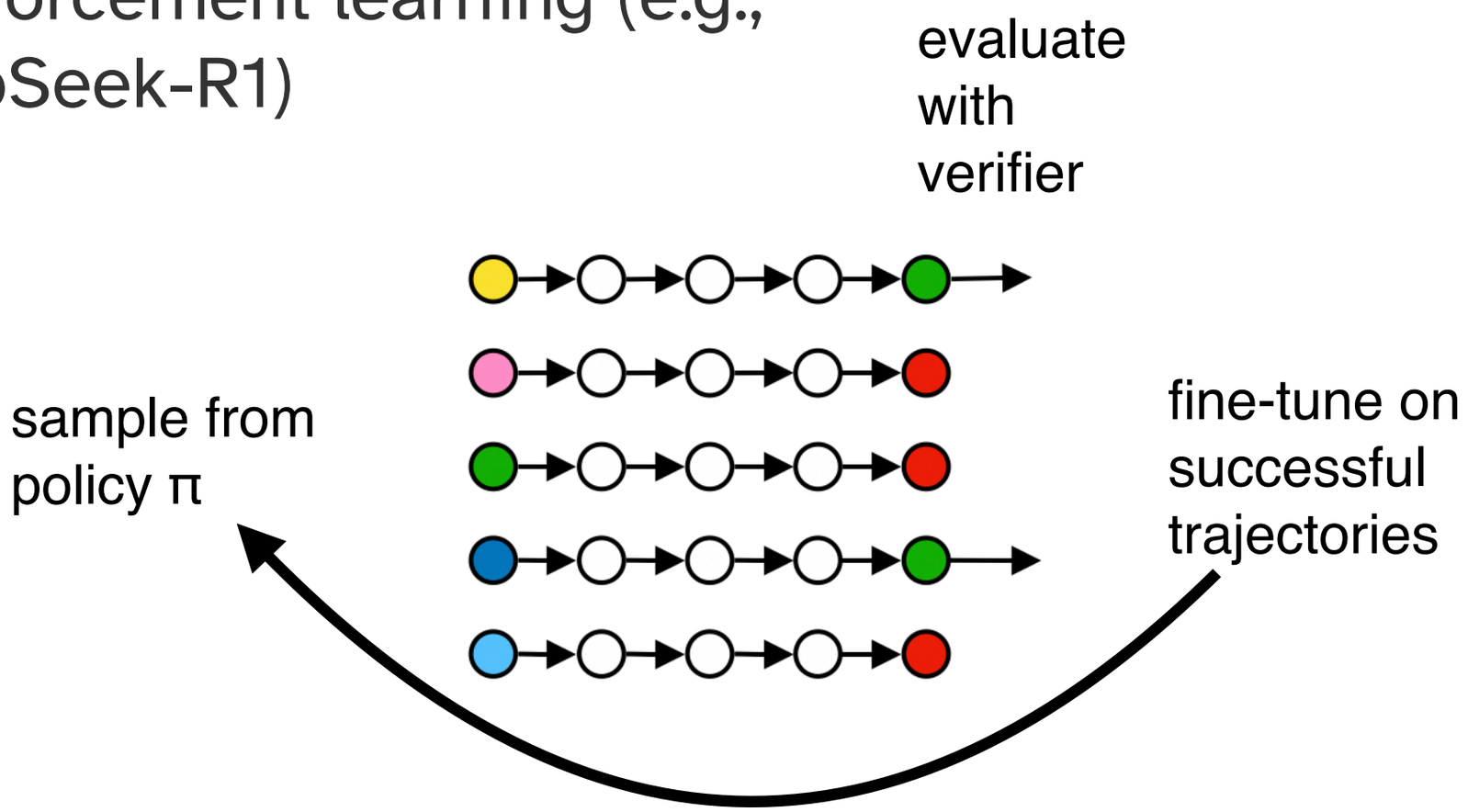
Types of reasoning models

- Training methods
  - Off-policy learning (e.g., s1, Muennighoff et al. 2025)



Types of reasoning models

- Training methods
  - Off-policy learning (e.g., s1, Muennighoff et al. 2025)
  - Reinforcement learning (e.g., DeepSeek-R1)



Learning to traverse the reasoning tree

Reinforcement learning enables learning to search over the reasoning tree up and down, including strategies like backtracking

Question: If $a > 1$, then the sum of the real solutions of $\sqrt{a - \sqrt{a+x}} = x$ is equal to

Response: <think>

To solve the equation $\sqrt{a - \sqrt{a+x}} = x$, let's start by squaring both $\cdots$

$\left(\sqrt{a - \sqrt{a+x}}\right)^2 = x^2 \implies a - \sqrt{a+x} = x^2$.

Rearrange to isolate the inner square root term:

$(a - x^2)^2 = a + x \implies a^2 - 2ax^2 + (x^2)^2 = a + x \implies x^4 - 2ax^2 - x + (a^2 - a) = 0$

$\cdots$

Wait, wait. Wait. That's an aha moment I can flag here.

Let's reevaluate this step-by-step to identify if the correct sum can be $\cdots$

We started with the equation:

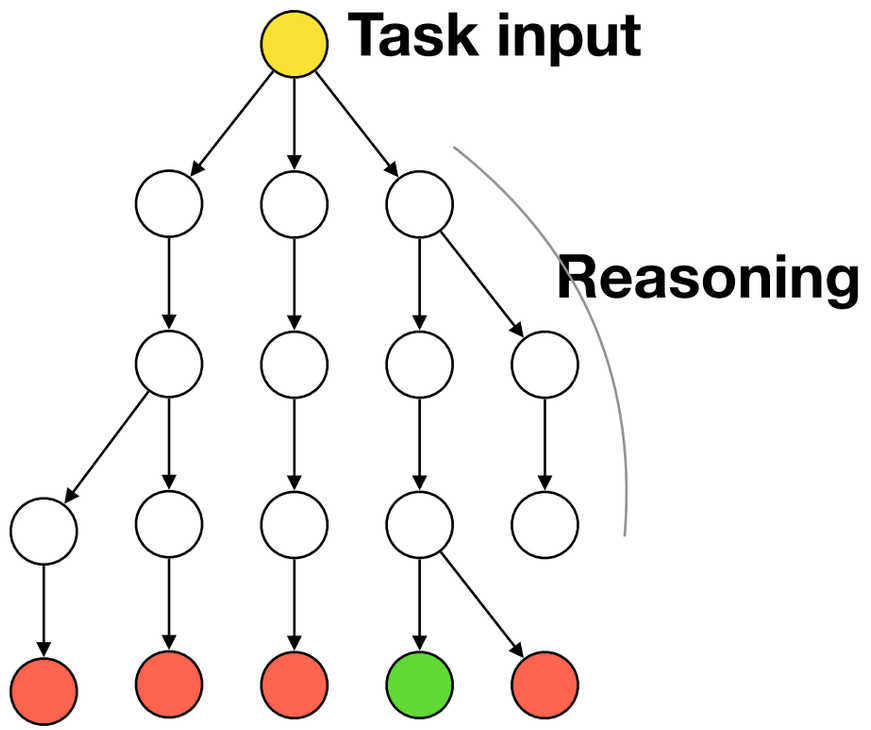
$\sqrt{a - \sqrt{a+x}} = x$

First, let's square both sides:

$a - \sqrt{a+x} = x^2 \implies \sqrt{a+x} = a - x^2$

Next, I could square both sides again, treating the equation: $\cdots$

$\cdots$



Learning to traverse the reasoning tree

Reinforcement learning enables learning to search over the reasoning tree up and down, including strategies like backtracking

Question: If $a > 1$, then the sum of the real solutions of $\sqrt{a - \sqrt{a + x}} = x$ is equal to

Response: <think>

To solve the equation $\sqrt{a - \sqrt{a + x}} = x$, let's start by squaring both $\cdots$

$\left(\sqrt{a - \sqrt{a + x}}\right)^2 = x^2 \implies a - \sqrt{a + x} = x^2$.

Rearrange to isolate the inner square root term:

$(a - x^2)^2 = a + x \implies a^2 - 2ax^2 + (x^2)^2 = a + x \implies x^4 - 2ax^2 - x + (a^2 - a) = 0$

$\cdots$

Wait, wait. Wait. That's an aha moment I can flag here.

Let's reevaluate this step-by-step to identify if the correct sum can be $\cdots$

We started with the equation:

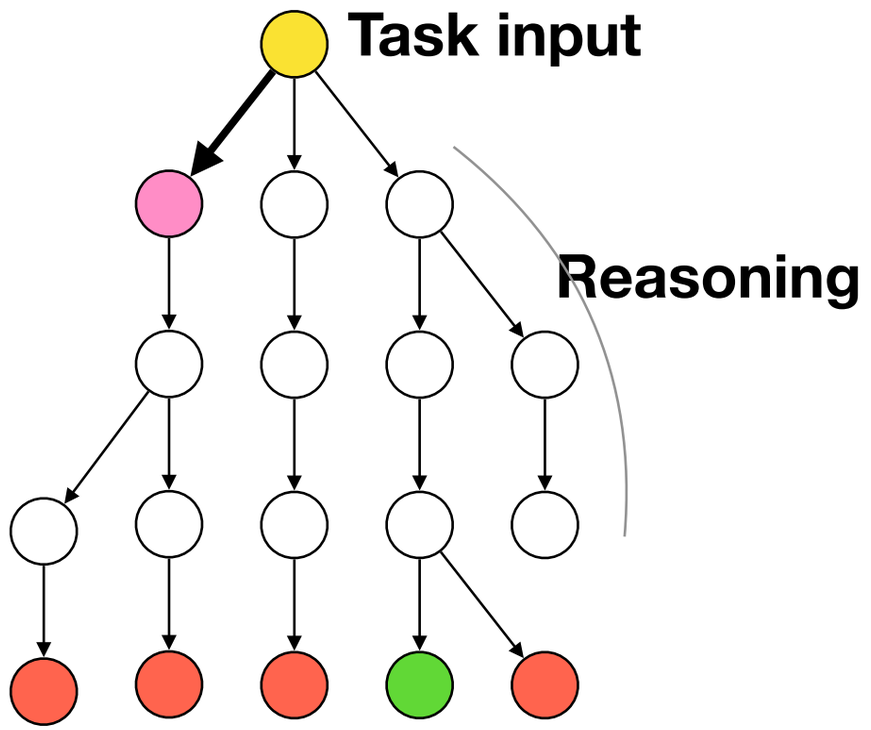
$\sqrt{a - \sqrt{a + x}} = x$

First, let's square both sides:

$a - \sqrt{a + x} = x^2 \implies \sqrt{a + x} = a - x^2$

Next, I could square both sides again, treating the equation: $\cdots$

$\cdots$



Learning to traverse the reasoning tree

Reinforcement learning enables learning to search over the reasoning tree up and down, including strategies like backtracking

Question: If $a > 1$, then the sum of the real solutions of $\sqrt{a - \sqrt{a + x}} = x$ is equal to

Response: <think>

To solve the equation $\sqrt{a - \sqrt{a + x}} = x$, let's start by squaring both $\cdots$

$\left(\sqrt{a - \sqrt{a + x}}\right)^2 = x^2 \implies a - \sqrt{a + x} = x^2$.

Rearrange to isolate the inner square root term:

$(a - x^2)^2 = a + x \implies a^2 - 2ax^2 + (x^2)^2 = a + x \implies x^4 - 2ax^2 - x + (a^2 - a) = 0$

$\ldots$

Wait, wait. Wait. That's an aha moment I can flag here.

Let's reevaluate this step-by-step to identify if the correct sum can be $\cdots$

We started with the equation:

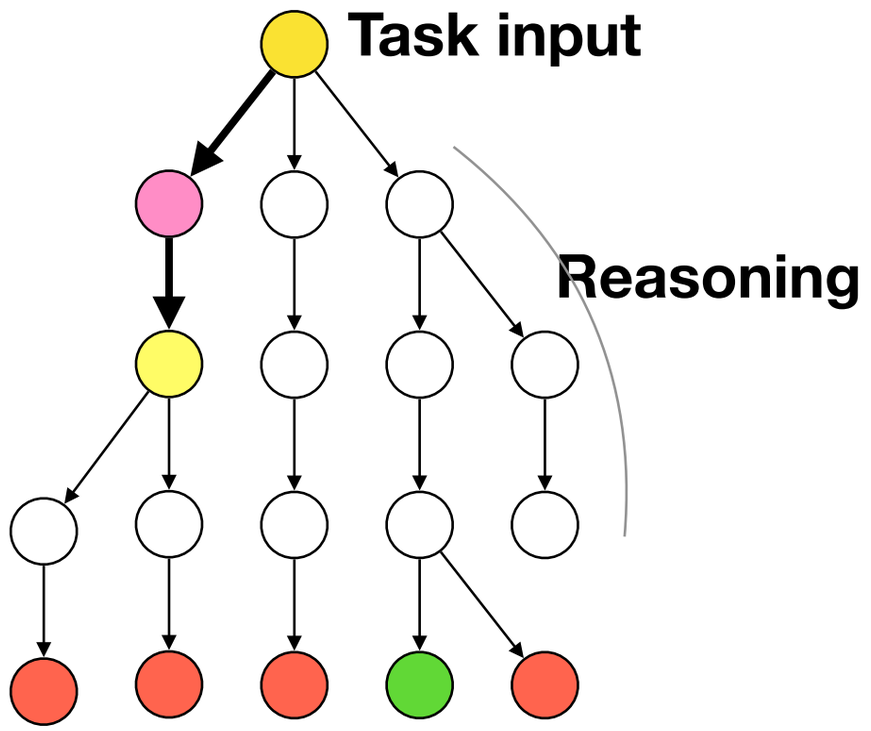
$\sqrt{a - \sqrt{a + x}} = x$

First, let's square both sides:

$a - \sqrt{a + x} = x^2 \implies \sqrt{a + x} = a - x^2$

Next, I could square both sides again, treating the equation: $\cdots$

$\ldots$



Learning to traverse the reasoning tree

Reinforcement learning enables learning to search over the reasoning tree up and down, including strategies like backtracking

Question: If $a > 1$, then the sum of the real solutions of $\sqrt{a - \sqrt{a+x}} = x$ is equal to

Response: <think>

To solve the equation $\sqrt{a - \sqrt{a+x}} = x$, let's start by squaring both $\cdots$

$\left(\sqrt{a - \sqrt{a+x}}\right)^2 = x^2 \implies a - \sqrt{a+x} = x^2$.

Rearrange to isolate the inner square root term:

$(a - x^2)^2 = a + x \implies a^2 - 2ax^2 + (x^2)^2 = a + x \implies x^4 - 2ax^2 - x + (a^2 - a) = 0$

$\cdots$

Wait, wait. Wait. That's an aha moment I can flag here.

Let's reevaluate this step-by-step to identify if the correct sum can be $\cdots$

We started with the equation:

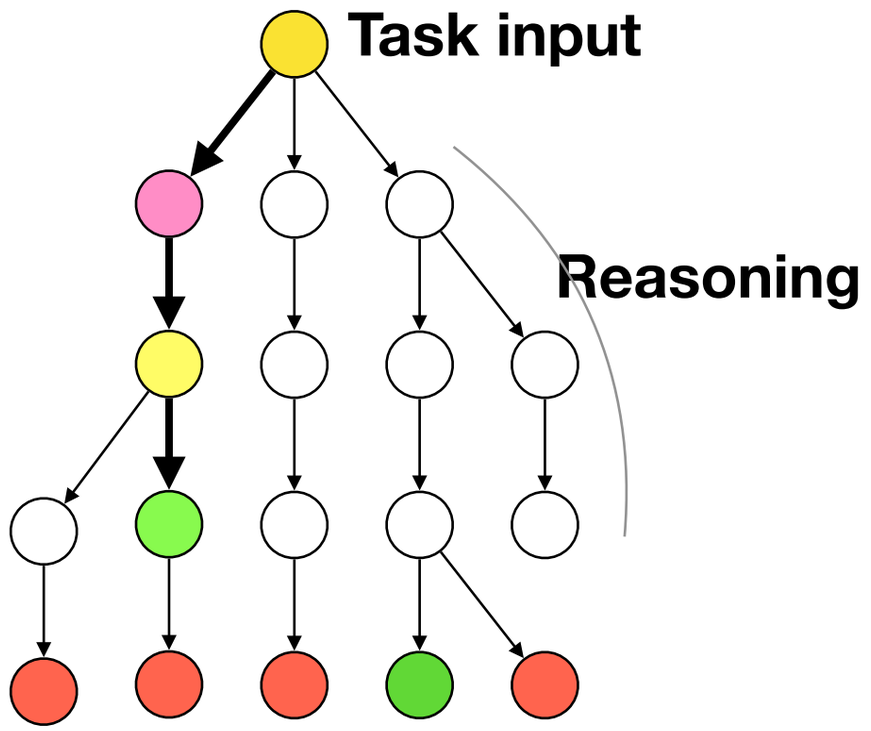
$\sqrt{a - \sqrt{a+x}} = x$

First, let's square both sides:

$a - \sqrt{a+x} = x^2 \implies \sqrt{a+x} = a - x^2$

Next, I could square both sides again, treating the equation: $\cdots$

$\cdots$



Learning to traverse the reasoning tree

Reinforcement learning enables learning to search over the reasoning tree up and down, including strategies like backtracking

Question: If $a > 1$, then the sum of the real solutions of $\sqrt{a - \sqrt{a+x}} = x$ is equal to

Response: <think>

To solve the equation $\sqrt{a - \sqrt{a+x}} = x$, let's start by squaring both $\cdots$

$\left(\sqrt{a - \sqrt{a+x}}\right)^2 = x^2 \implies a - \sqrt{a+x} = x^2$.

Rearrange to isolate the inner square root term:

$(a - x^2)^2 = a + x \implies a^2 - 2ax^2 + (x^2)^2 = a + x \implies x^4 - 2ax^2 - x + (a^2 - a) = 0$

$\cdots$

Wait, wait. Wait. That's an aha moment I can flag here.

Let's reevaluate this step-by-step to identify if the correct sum can be $\cdots$

We started with the equation:

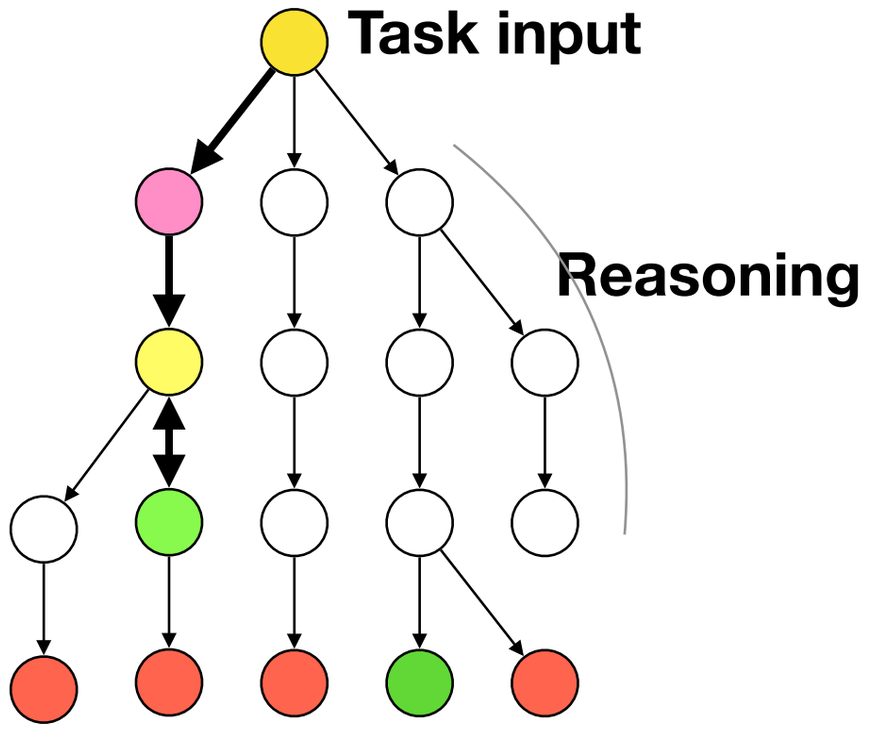
$\sqrt{a - \sqrt{a+x}} = x$

First, let's square both sides:

$a - \sqrt{a+x} = x^2 \implies \sqrt{a+x} = a - x^2$

Next, I could square both sides again, treating the equation: $\cdots$

$\cdots$



Learning to traverse the reasoning tree

Reinforcement learning enables learning to search over the reasoning tree up and down, including strategies like backtracking

Question: If $a > 1$, then the sum of the real solutions of $\sqrt{a - \sqrt{a+x}} = x$ is equal to

Response: <think>

To solve the equation $\sqrt{a - \sqrt{a+x}} = x$, let's start by squaring both $\cdots$

$\left(\sqrt{a - \sqrt{a+x}}\right)^2 = x^2 \implies a - \sqrt{a+x} = x^2$.

Rearrange to isolate the inner square root term:

$(a - x^2)^2 = a + x \implies a^2 - 2ax^2 + (x^2)^2 = a + x \implies x^4 - 2ax^2 - x + (a^2 - a) = 0$

$\ldots$

Wait, wait. Wait. That's an aha moment I can flag here.

Let's reevaluate this step-by-step to identify if the correct sum can be $\cdots$

We started with the equation:

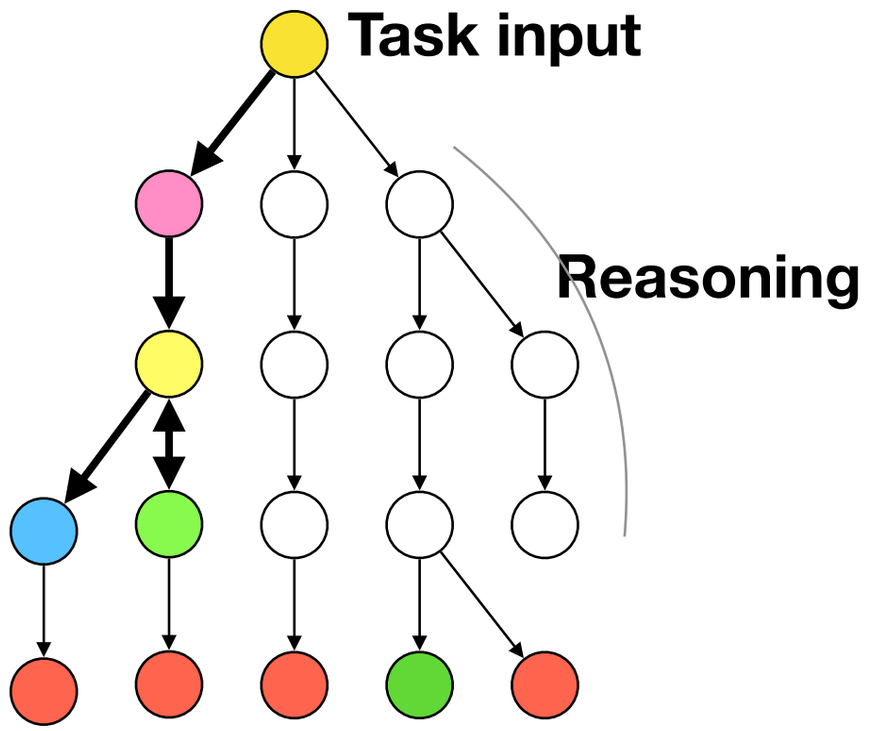
$\sqrt{a - \sqrt{a+x}} = x$

First, let's square both sides:

$a - \sqrt{a+x} = x^2 \implies \sqrt{a+x} = a - x^2$

Next, I could square both sides again, treating the equation: $\cdots$

$\ldots$



Learning to traverse the reasoning tree

Reinforcement learning enables learning to search over the reasoning tree up and down, including strategies like backtracking

Question: If $a > 1$, then the sum of the real solutions of $\sqrt{a - \sqrt{a+x}} = x$ is equal to

Response: <think>

To solve the equation $\sqrt{a - \sqrt{a+x}} = x$, let's start by squaring both $\cdots$

$\left(\sqrt{a - \sqrt{a+x}}\right)^2 = x^2 \implies a - \sqrt{a+x} = x^2$.

Rearrange to isolate the inner square root term:

$(a - x^2)^2 = a + x \implies a^2 - 2ax^2 + (x^2)^2 = a + x \implies x^4 - 2ax^2 - x + (a^2 - a) = 0$

$\ldots$

Wait, wait. Wait. That's an aha moment I can flag here.

Let's reevaluate this step-by-step to identify if the correct sum can be $\cdots$

We started with the equation:

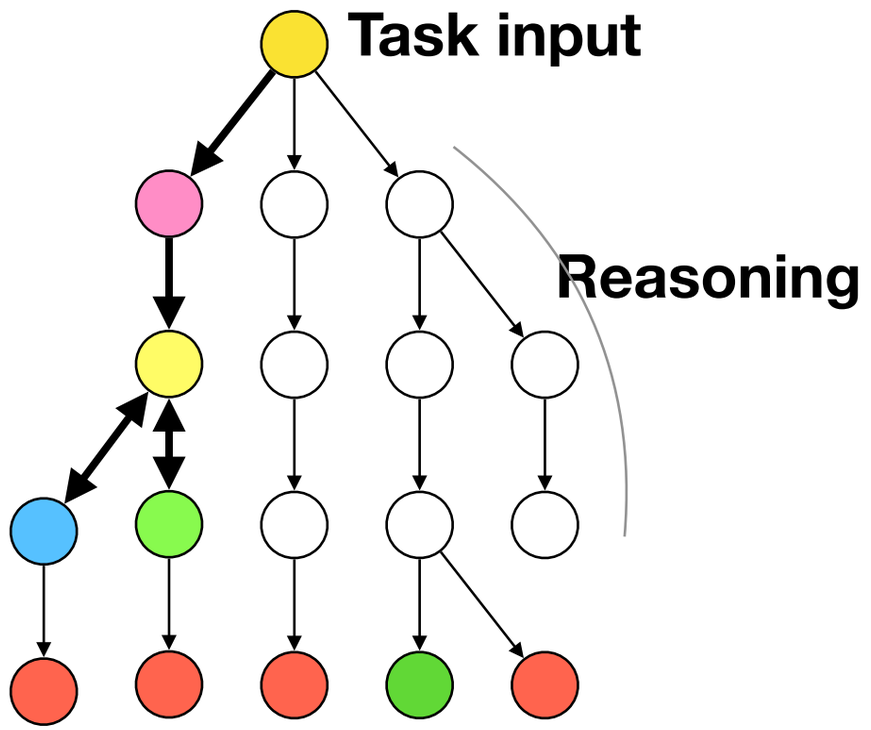
$\sqrt{a - \sqrt{a+x}} = x$

First, let's square both sides:

$a - \sqrt{a+x} = x^2 \implies \sqrt{a+x} = a - x^2$

Next, I could square both sides again, treating the equation: $\cdots$

$\ldots$



Learning to traverse the reasoning tree

Reinforcement learning enables learning to search over the reasoning tree up and down, including strategies like backtracking

Question: If $a > 1$, then the sum of the real solutions of $\sqrt{a - \sqrt{a+x}} = x$ is equal to

Response: <think>

To solve the equation $\sqrt{a - \sqrt{a+x}} = x$, let's start by squaring both $\cdots$

$\left(\sqrt{a - \sqrt{a+x}}\right)^2 = x^2 \implies a - \sqrt{a+x} = x^2$.

Rearrange to isolate the inner square root term:

$(a - x^2)^2 = a + x \implies a^2 - 2ax^2 + (x^2)^2 = a + x \implies x^4 - 2ax^2 - x + (a^2 - a) = 0$

$\cdots$

Wait, wait. Wait. That's an aha moment I can flag here.

Let's reevaluate this step-by-step to identify if the correct sum can be $\cdots$

We started with the equation:

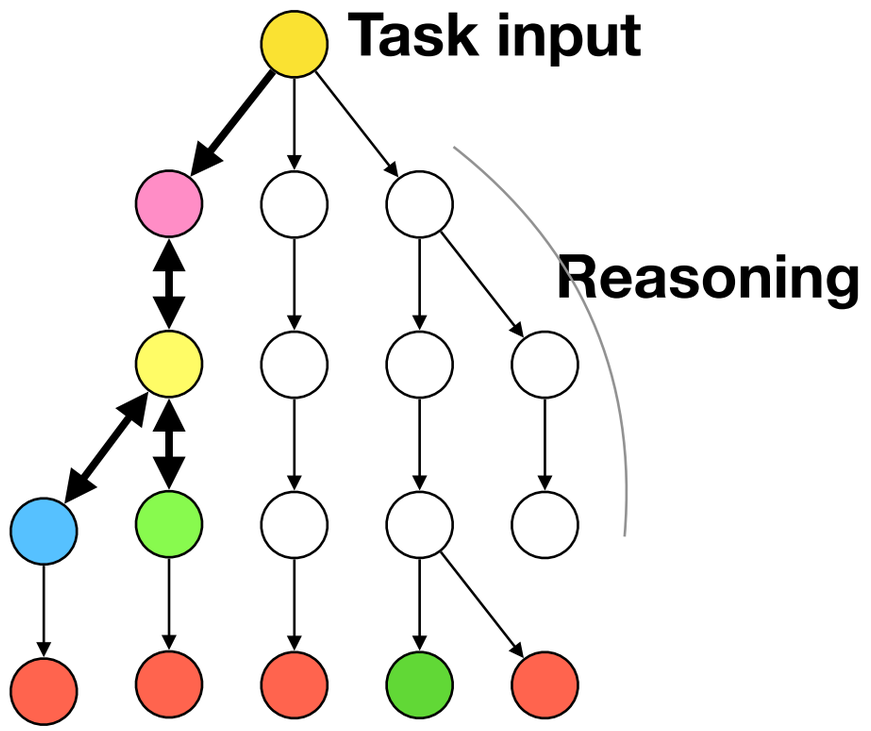
$\sqrt{a - \sqrt{a+x}} = x$

First, let's square both sides:

$a - \sqrt{a+x} = x^2 \implies \sqrt{a+x} = a - x^2$

Next, I could square both sides again, treating the equation: $\cdots$

$\cdots$



Learning to traverse the reasoning tree

Reinforcement learning enables learning to search over the reasoning tree up and down, including strategies like backtracking

Question: If $a > 1$, then the sum of the real solutions of $\sqrt{a - \sqrt{a+x}} = x$ is equal to

Response: <think>

To solve the equation $\sqrt{a - \sqrt{a+x}} = x$, let's start by squaring both $\cdots$

$\left(\sqrt{a - \sqrt{a+x}}\right)^2 = x^2 \implies a - \sqrt{a+x} = x^2$.

Rearrange to isolate the inner square root term:

$(a - x^2)^2 = a + x \implies a^2 - 2ax^2 + (x^2)^2 = a + x \implies x^4 - 2ax^2 - x + (a^2 - a) = 0$

$\cdots$

Wait, wait. Wait. That's an aha moment I can flag here.

Let's reevaluate this step-by-step to identify if the correct sum can be $\cdots$

We started with the equation:

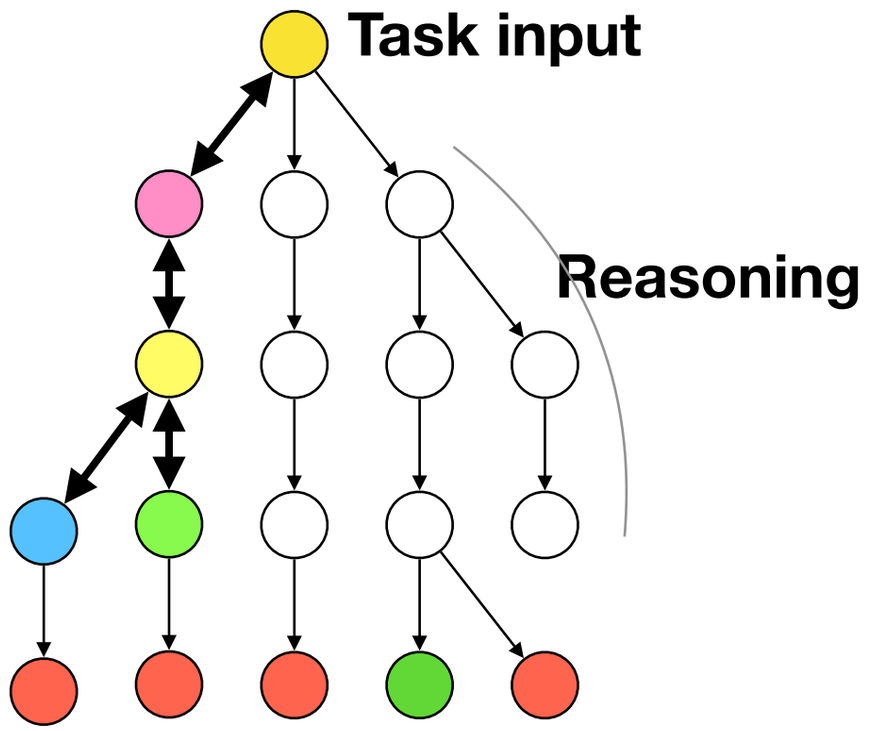
$\sqrt{a - \sqrt{a+x}} = x$

First, let's square both sides:

$a - \sqrt{a+x} = x^2 \implies \sqrt{a+x} = a - x^2$

Next, I could square both sides again, treating the equation: $\cdots$

$\cdots$



Learning to traverse the reasoning tree

Reinforcement learning enables learning to search over the reasoning tree up and down, including strategies like backtracking

Question: If $a > 1$, then the sum of the real solutions of $\sqrt{a - \sqrt{a + x}} = x$ is equal to

Response: <think>

To solve the equation $\sqrt{a - \sqrt{a + x}} = x$, let's start by squaring both $\cdots$

$\left(\sqrt{a - \sqrt{a + x}}\right)^2 = x^2 \implies a - \sqrt{a + x} = x^2$.

Rearrange to isolate the inner square root term:

$(a - x^2)^2 = a + x \implies a^2 - 2ax^2 + (x^2)^2 = a + x \implies x^4 - 2ax^2 - x + (a^2 - a) = 0$

$\cdots$

Wait, wait. Wait. That's an aha moment I can flag here.

Let's reevaluate this step-by-step to identify if the correct sum can be $\cdots$

We started with the equation:

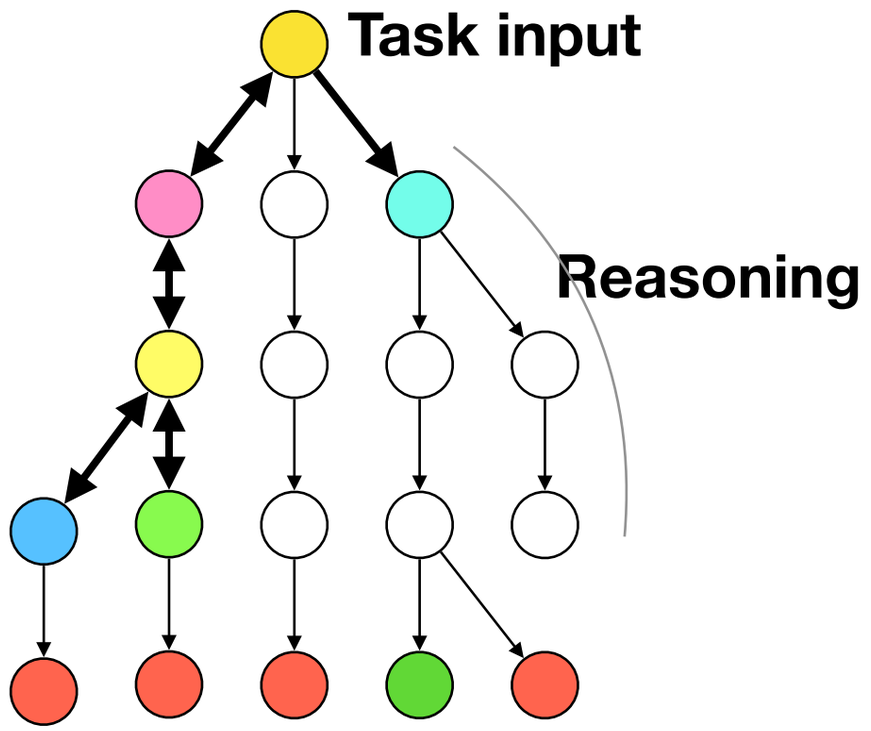
$\sqrt{a - \sqrt{a + x}} = x$

First, let's square both sides:

$a - \sqrt{a + x} = x^2 \implies \sqrt{a + x} = a - x^2$

Next, I could square both sides again, treating the equation: $\cdots$

$\cdots$



Learning to traverse the reasoning tree

Reinforcement learning enables learning to search over the reasoning tree up and down, including strategies like backtracking

Question: If $a > 1$, then the sum of the real solutions of $\sqrt{a - \sqrt{a + x}} = x$ is equal to

Response: <think>

To solve the equation $\sqrt{a - \sqrt{a + x}} = x$, let's start by squaring both $\cdots$

$\left(\sqrt{a - \sqrt{a + x}}\right)^2 = x^2 \implies a - \sqrt{a + x} = x^2$.

Rearrange to isolate the inner square root term:

$(a - x^2)^2 = a + x \implies a^2 - 2ax^2 + (x^2)^2 = a + x \implies x^4 - 2ax^2 - x + (a^2 - a) = 0$

$\cdots$

Wait, wait. Wait. That's an aha moment I can flag here.

Let's reevaluate this step-by-step to identify if the correct sum can be $\cdots$

We started with the equation:

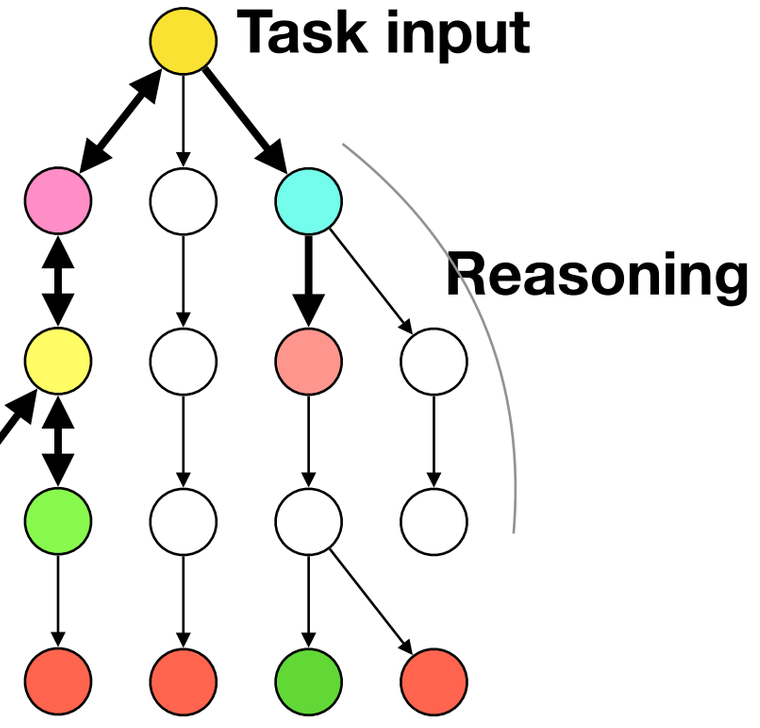
$\sqrt{a - \sqrt{a + x}} = x$

First, let's square both sides:

$a - \sqrt{a + x} = x^2 \implies \sqrt{a + x} = a - x^2$

Next, I could square both sides again, treating the equation: $\cdots$

$\cdots$



Learning to traverse the reasoning tree

Reinforcement learning enables learning to search over the reasoning tree up and down, including strategies like backtracking

Question: If $a > 1$, then the sum of the real solutions of $\sqrt{a - \sqrt{a + x}} = x$ is equal to

Response: <think>

To solve the equation $\sqrt{a - \sqrt{a + x}} = x$, let's start by squaring both $\cdots$

$\left(\sqrt{a - \sqrt{a + x}}\right)^2 = x^2 \implies a - \sqrt{a + x} = x^2$.

Rearrange to isolate the inner square root term:

$(a - x^2)^2 = a + x \implies a^2 - 2ax^2 + (x^2)^2 = a + x \implies x^4 - 2ax^2 - x + (a^2 - a) = 0$

$\cdots$

Wait, wait. Wait. That's an aha moment I can flag here.

Let's reevaluate this step-by-step to identify if the correct sum can be $\cdots$

We started with the equation:

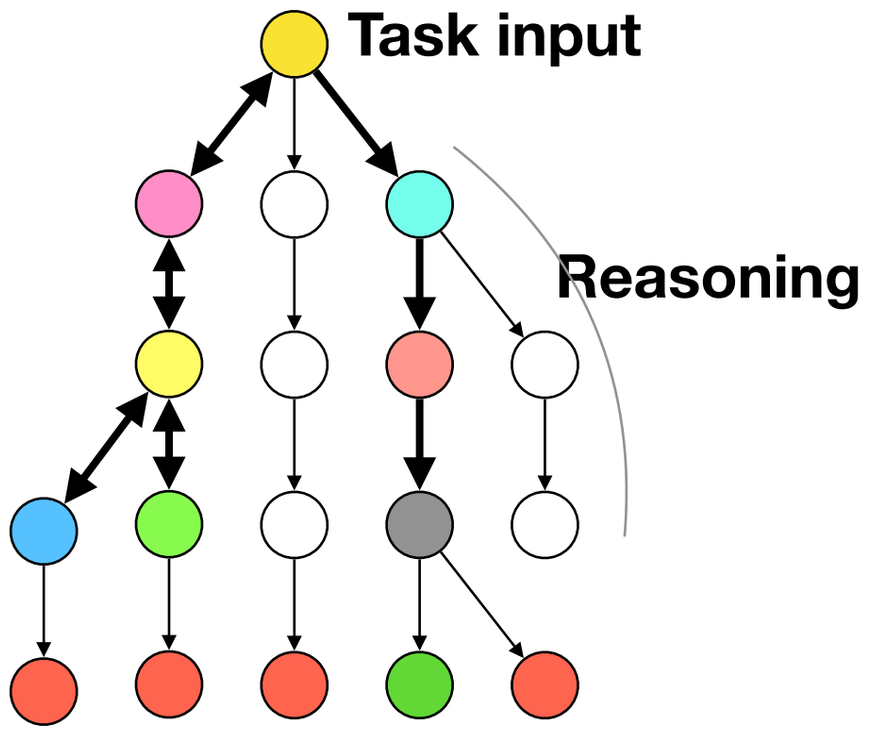
$\sqrt{a - \sqrt{a + x}} = x$

First, let's square both sides:

$a - \sqrt{a + x} = x^2 \implies \sqrt{a + x} = a - x^2$

Next, I could square both sides again, treating the equation: $\cdots$

$\cdots$



Learning to traverse the reasoning tree

Reinforcement learning enables learning to search over the reasoning tree up and down, including strategies like backtracking

Question: If $a > 1$, then the sum of the real solutions of $\sqrt{a - \sqrt{a+x}} = x$ is equal to

Response: <think>

To solve the equation $\sqrt{a - \sqrt{a+x}} = x$, let's start by squaring both $\cdots$

$\left(\sqrt{a - \sqrt{a+x}}\right)^2 = x^2 \implies a - \sqrt{a+x} = x^2$.

Rearrange to isolate the inner square root term:

$(a - x^2)^2 = a + x \implies a^2 - 2ax^2 + (x^2)^2 = a + x \implies x^4 - 2ax^2 - x + (a^2 - a) = 0$

$\cdots$

Wait, wait. Wait. That's an aha moment I can flag here.

Let's reevaluate this step-by-step to identify if the correct sum can be $\cdots$

We started with the equation:

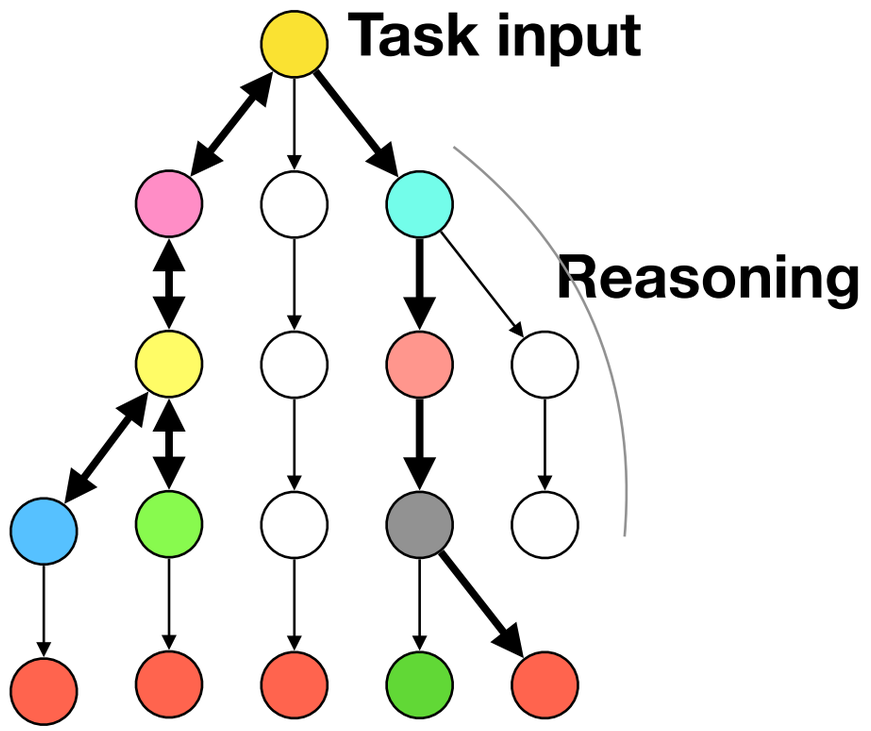
$\sqrt{a - \sqrt{a+x}} = x$

First, let's square both sides:

$a - \sqrt{a+x} = x^2 \implies \sqrt{a+x} = a - x^2$

Next, I could square both sides again, treating the equation: $\cdots$

$\cdots$



Learning to traverse the reasoning tree

Reinforcement learning enables learning to search over the reasoning tree up and down, including strategies like backtracking

Question: If $a > 1$, then the sum of the real solutions of $\sqrt{a - \sqrt{a + x}} = x$ is equal to

Response: <think>

To solve the equation $\sqrt{a - \sqrt{a + x}} = x$, let's start by squaring both $\cdots$

$\left(\sqrt{a - \sqrt{a + x}}\right)^2 = x^2 \implies a - \sqrt{a + x} = x^2$.

Rearrange to isolate the inner square root term:

$(a - x^2)^2 = a + x \implies a^2 - 2ax^2 + (x^2)^2 = a + x \implies x^4 - 2ax^2 - x + (a^2 - a) = 0$

$\ldots$

Wait, wait. Wait. That's an aha moment I can flag here.

Let's reevaluate this step-by-step to identify if the correct sum can be $\cdots$

We started with the equation:

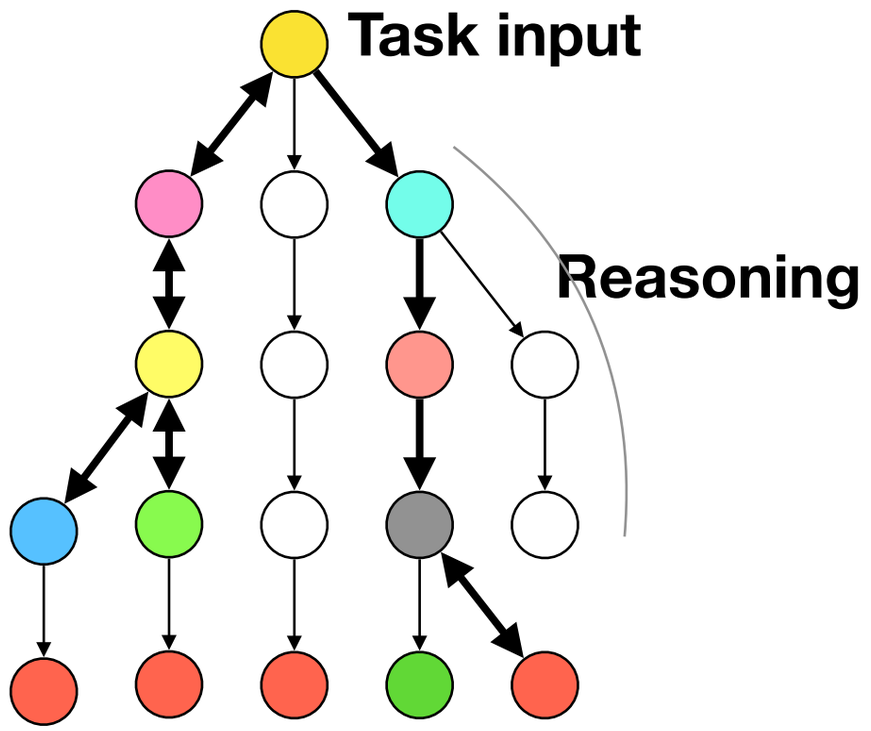
$\sqrt{a - \sqrt{a + x}} = x$

First, let's square both sides:

$a - \sqrt{a + x} = x^2 \implies \sqrt{a + x} = a - x^2$

Next, I could square both sides again, treating the equation: $\cdots$

$\ldots$



Learning to traverse the reasoning tree

Reinforcement learning enables learning to search over the reasoning tree up and down, including strategies like backtracking

Question: If $a > 1$, then the sum of the real solutions of $\sqrt{a - \sqrt{a + x}} = x$ is equal to

Response: <think>

To solve the equation $\sqrt{a - \sqrt{a + x}} = x$, let's start by squaring both $\cdots$

$\left(\sqrt{a - \sqrt{a + x}}\right)^2 = x^2 \implies a - \sqrt{a + x} = x^2$.

Rearrange to isolate the inner square root term:

$(a - x^2)^2 = a + x \implies a^2 - 2ax^2 + (x^2)^2 = a + x \implies x^4 - 2ax^2 - x + (a^2 - a) = 0$

$\ldots$

Wait, wait. Wait. That's an aha moment I can flag here.

Let's reevaluate this step-by-step to identify if the correct sum can be $\cdots$

We started with the equation:

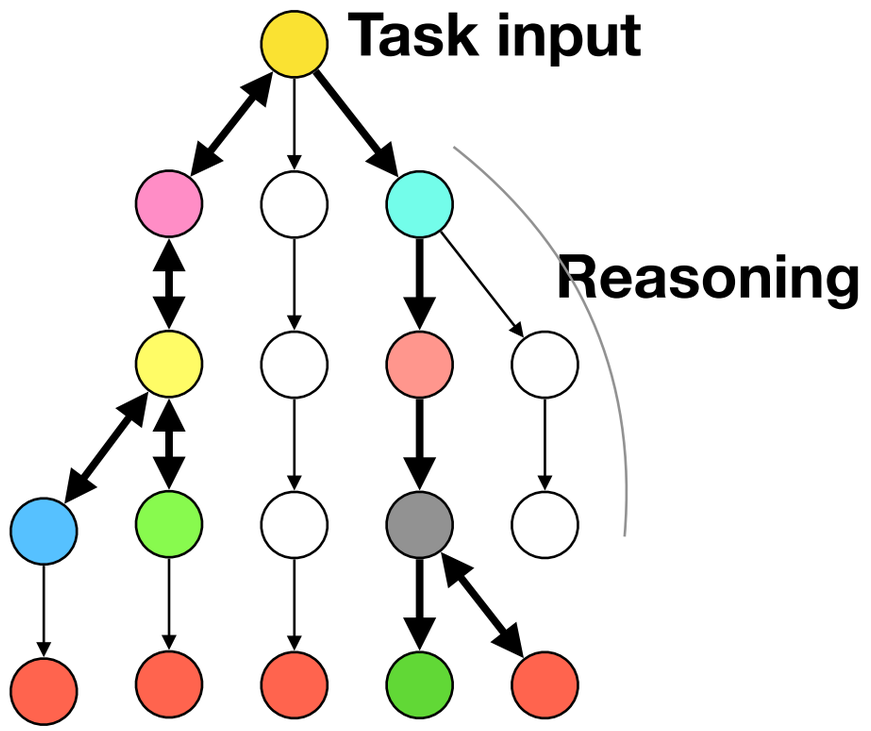
$\sqrt{a - \sqrt{a + x}} = x$

First, let's square both sides:

$a - \sqrt{a + x} = x^2 \implies \sqrt{a + x} = a - x^2$

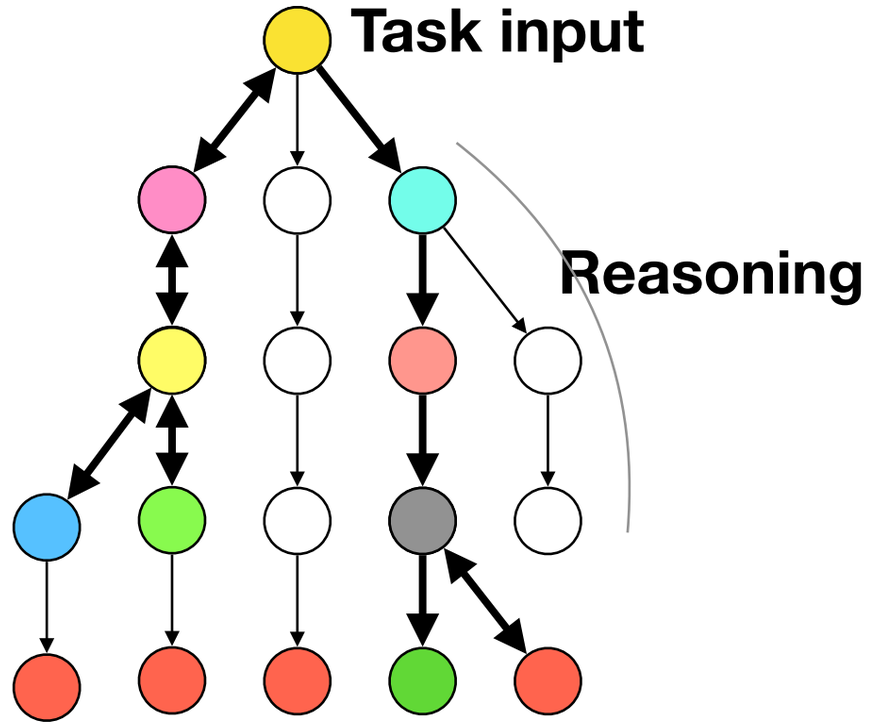
Next, I could square both sides again, treating the equation: $\cdots$

$\ldots$



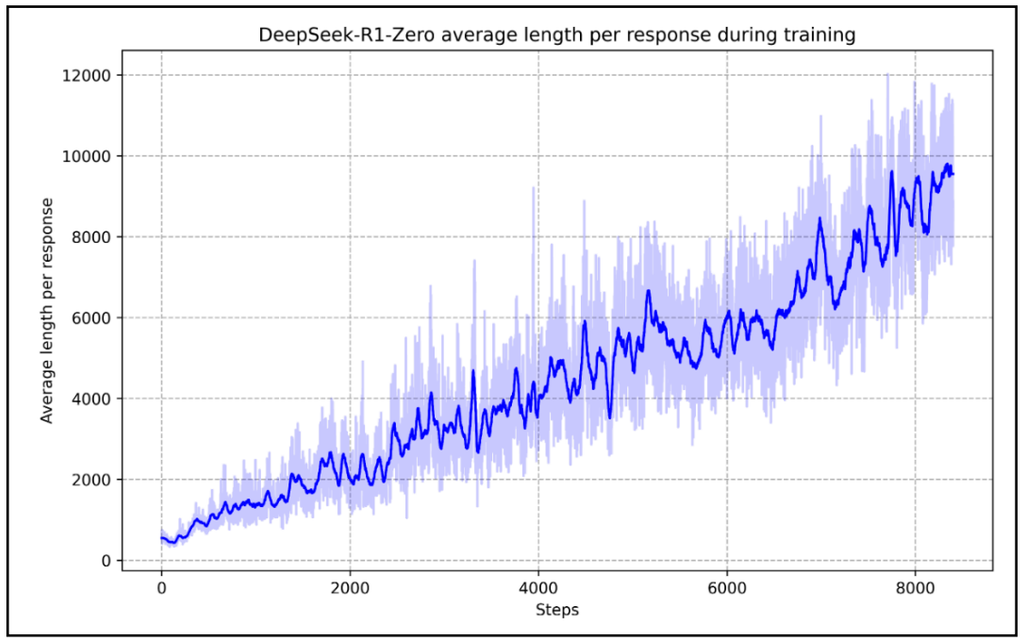
Learning to traverse the reasoning tree

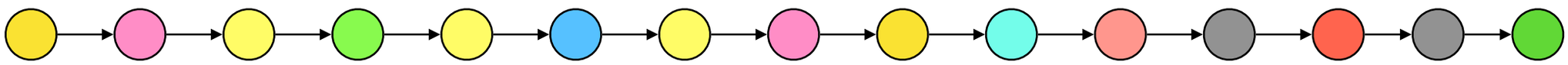
Reinforcement learning enables learning to search over the reasoning tree up and down, including strategies like backtracking



Learning to traverse the reasoning tree

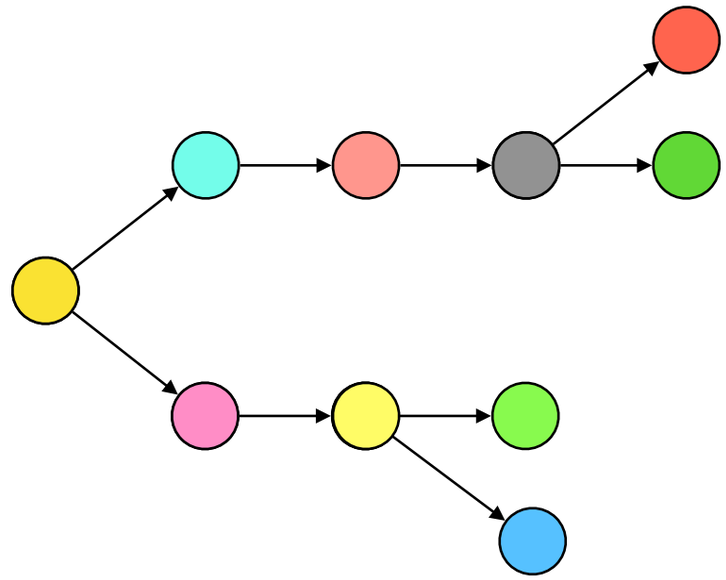
- RL incentivizes increasingly lengthy reasoning chains
  - Stresses context window of model architecture
  - Increases computational complexity of attention
- And the problem is inherently hierarchical!





Instead: adaptive parallel reasoning

- Parallel traversal of reasoning tree
- Language model is trained to distribute its inference-time compute across parallel and serial operations
- Two key components
  - Augmenting model output space with fork and join operations
  - Training with reinforcement learning to discover optimal inference structures

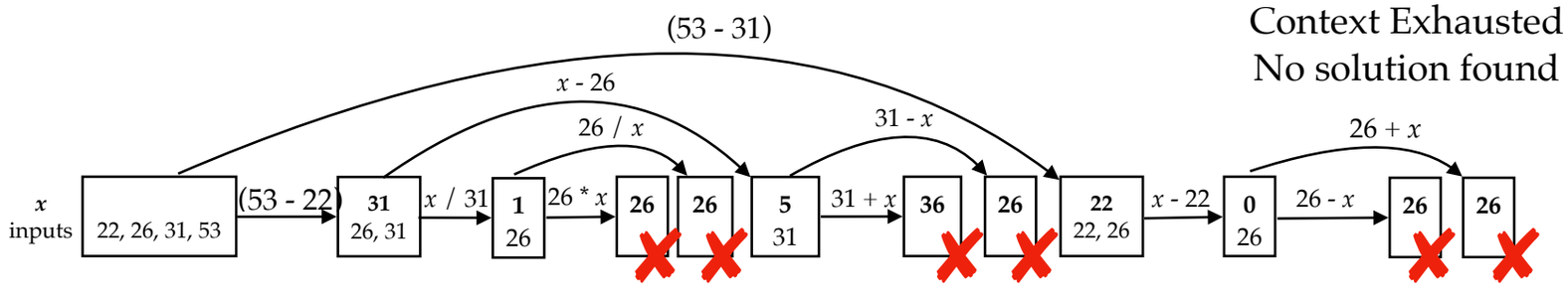


Fork and join

- Countdown task

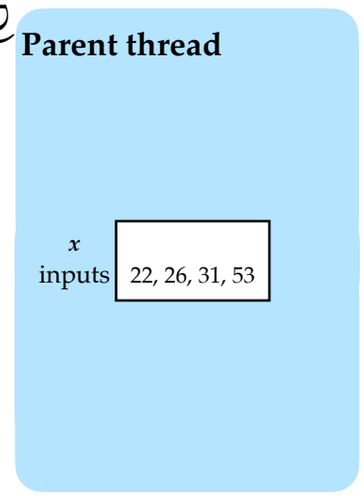
Target: 27
Inputs: 22, 26, 31, 53

SoS+ (Gandhi et al. 2024)



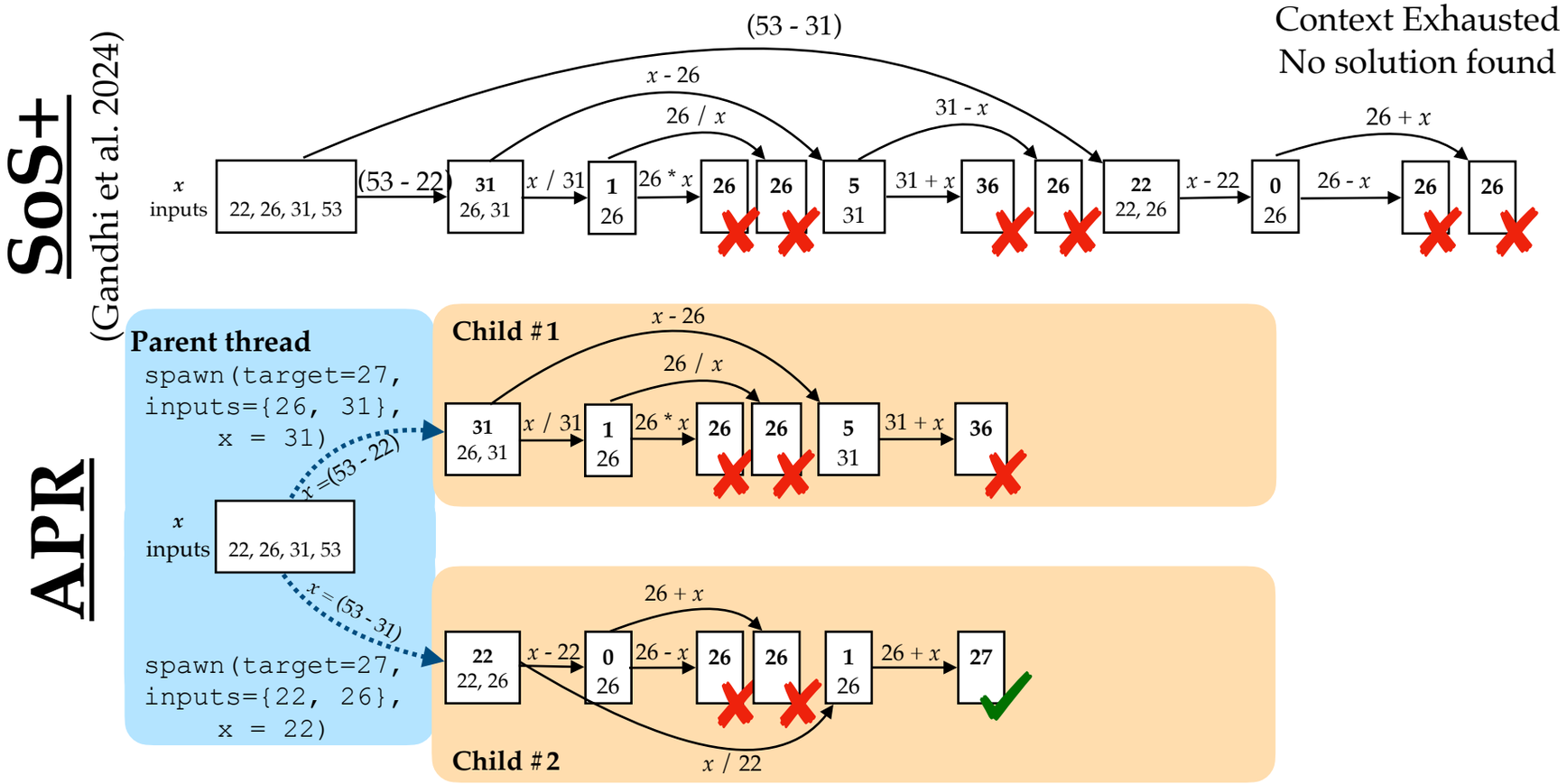
SoS+ Result:
Context Exhausted
No solution found

APR



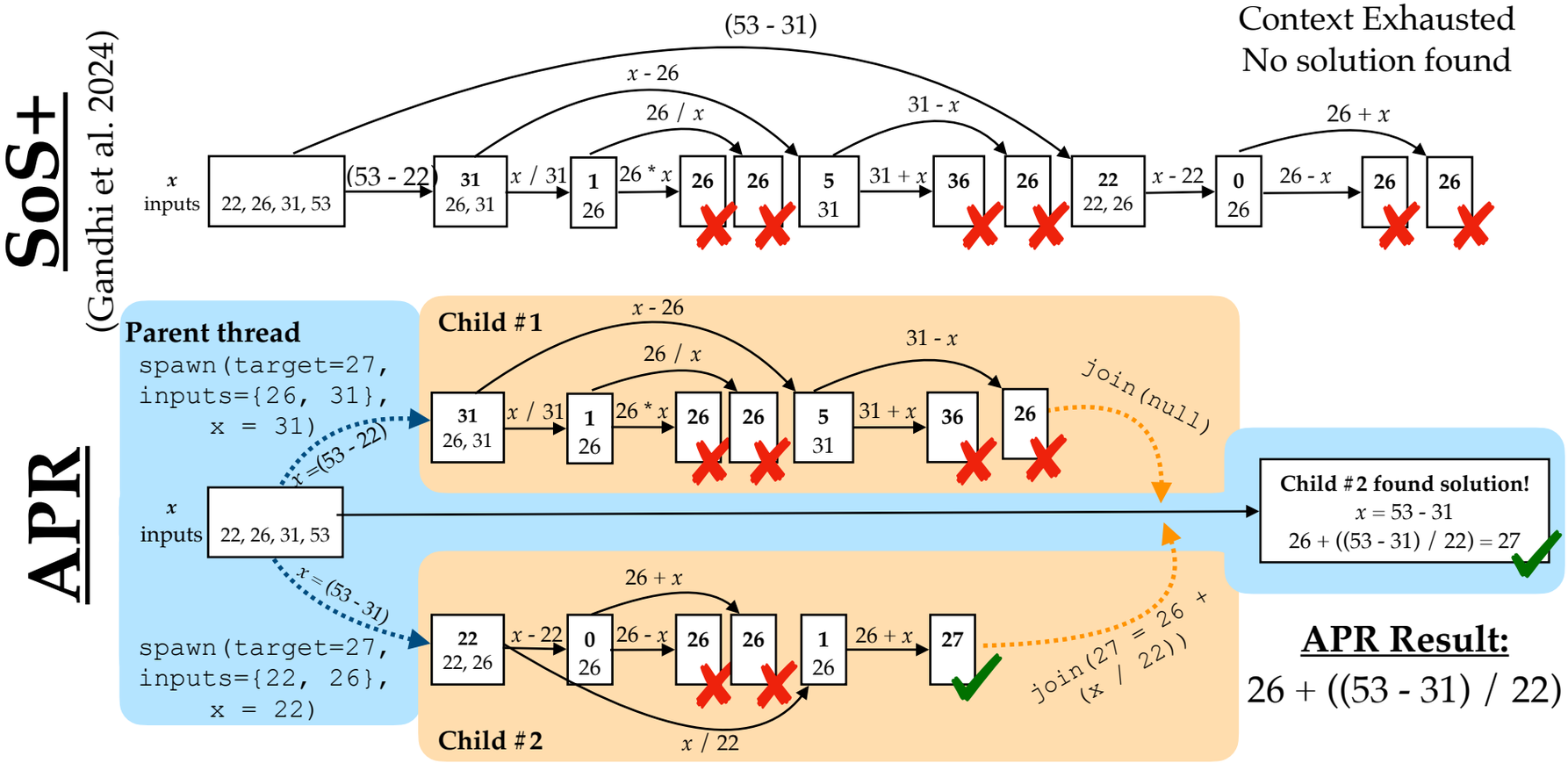
Fork and join

- Countdown task
- Fork
  - Send unique context to each child thread
- Child operates only on this context, reducing stress on context limits and filtering information



Fork and join

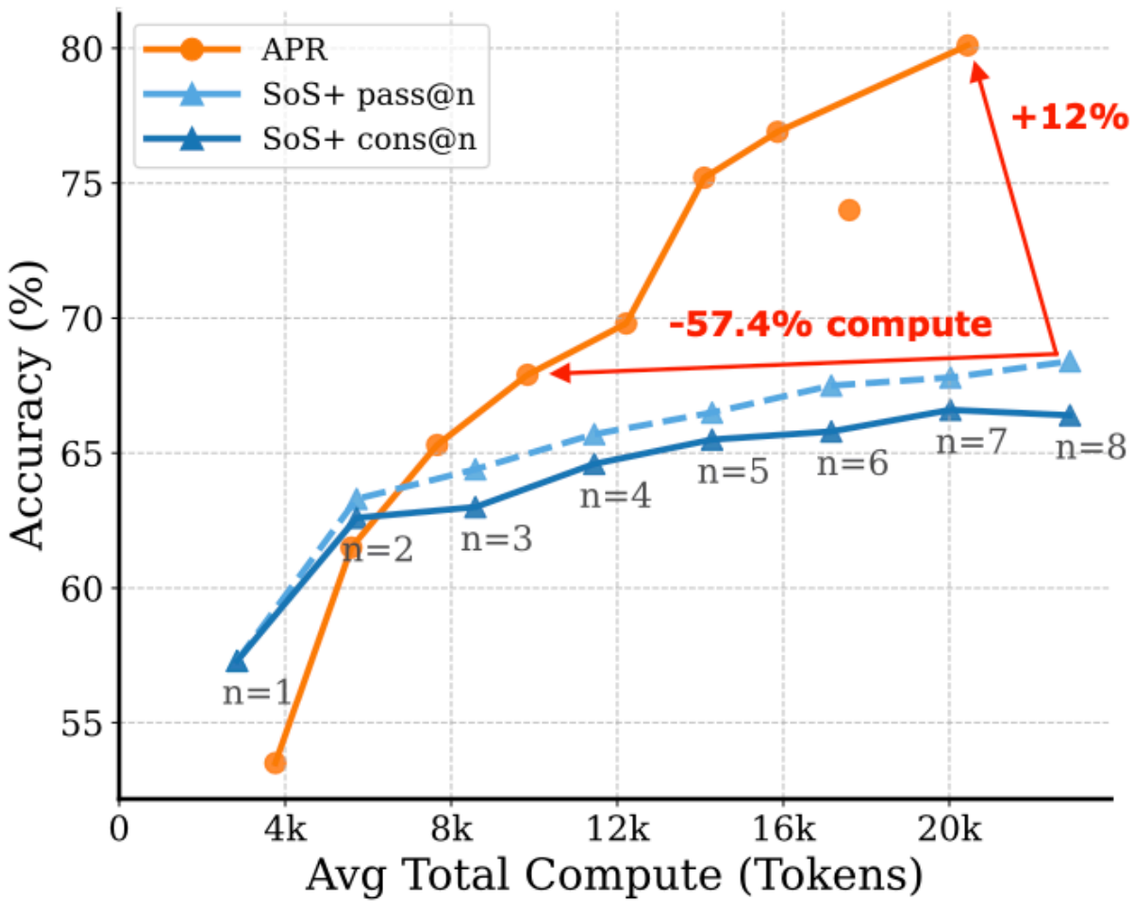
- Countdown task
- Fork
- Join
  - Child thread returns only its result, not its reasoning chain
- Parent thread synthesizes child results



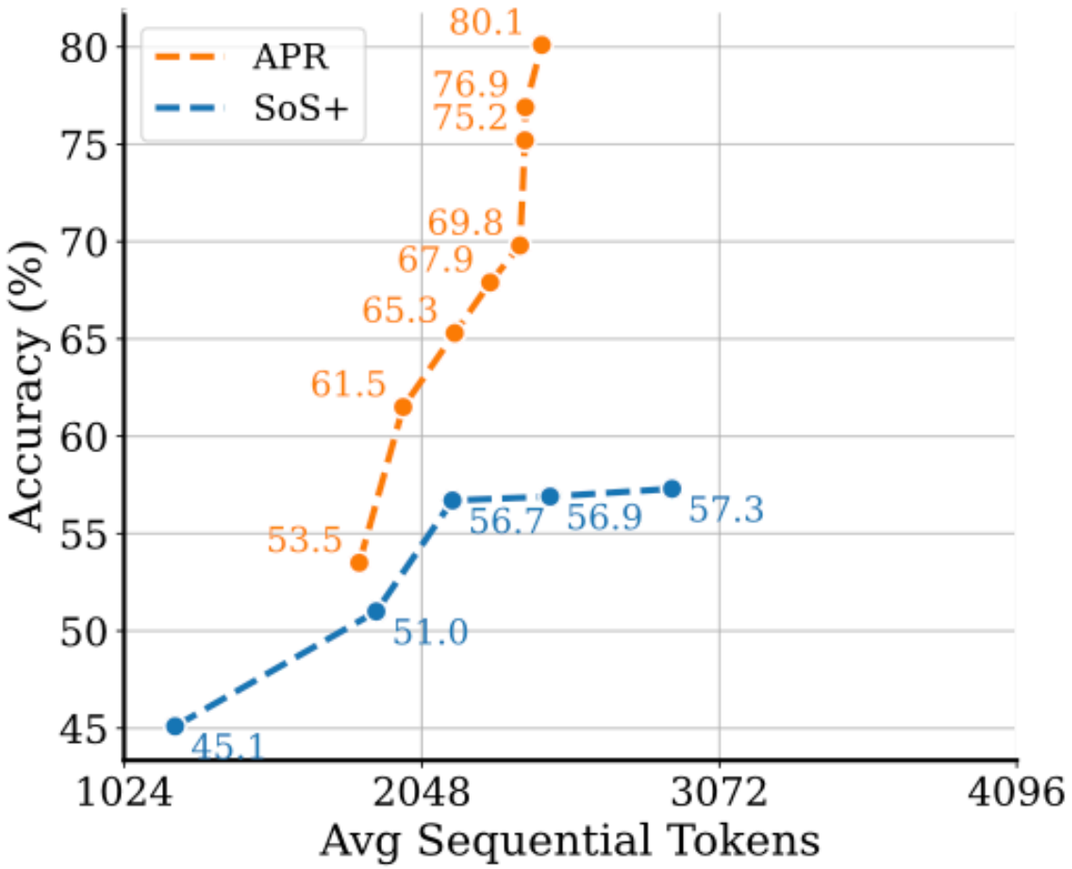
APR is token-efficient

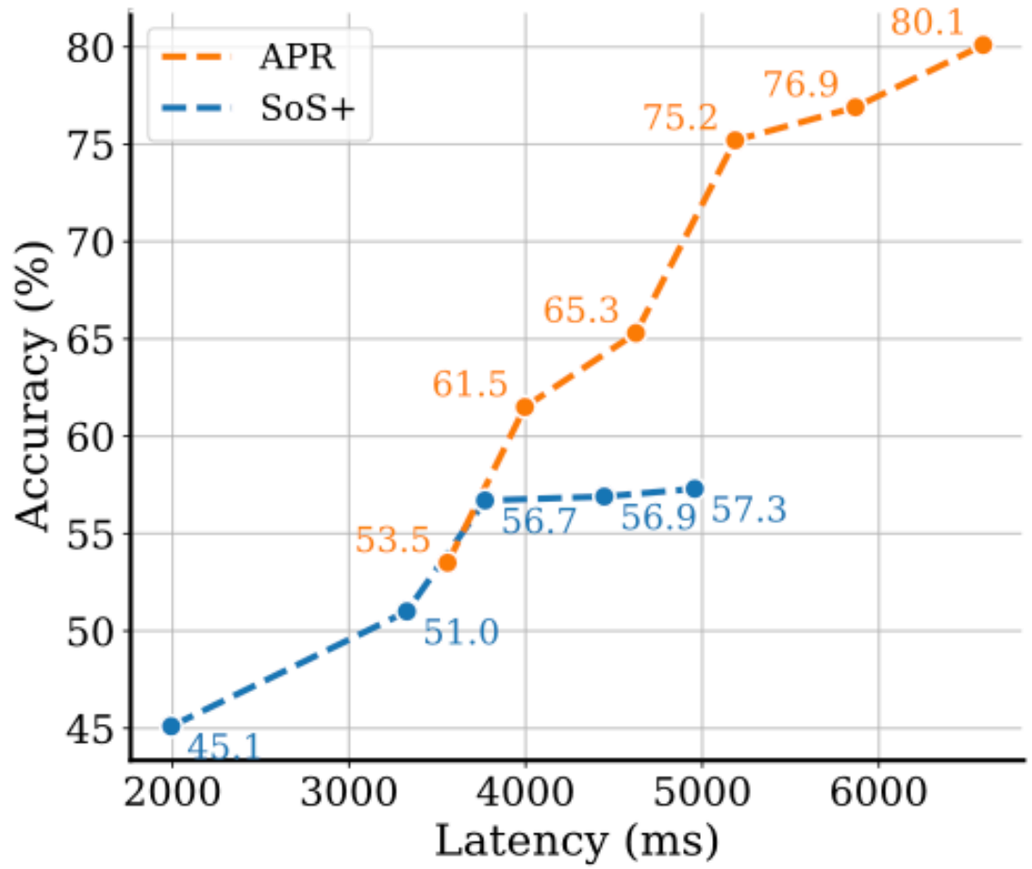
- Setup
  - Fixed window size of 4,096 tokens
  - Baseline: SoS with consensus

- Results
  - APR has equivalent performance of best SoS (pass@8) with less than half the compute
  - APR significantly improves over SoS pass@8 (68.4 to 80.1% accuracy) with fewer tokens (24k vs. 20k)



APR is realtime-efficient





Language Agents

Where do you see this going?

What would you want from a language agent?