

World Model Reasoning for Open-World Agents

Manling Li
Northwestern University
Amazon Scholar
manling.li@northwestern.edu



NORTHWESTERN
UNIVERSITY



MLL LAB
Machine Learning and Language

How Models Learn (and Fail to Learn) the Physical World



*A robot arm is placing
the banana on top of
the coaster*



What we expect



banana → monkey → jump

Output from Wan2.2

Reasoning

what we want

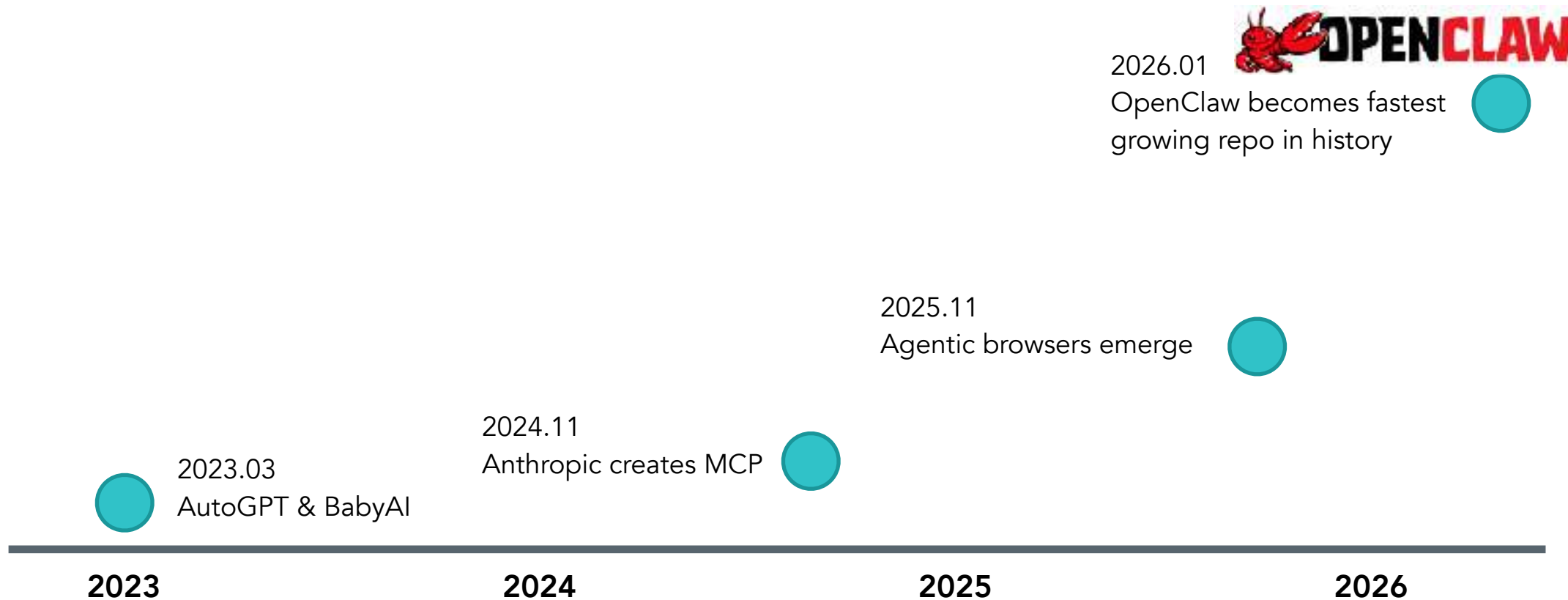
??

Memorization

what we truly learn

The Age of Agents

Capability



But Agents can fail **catastrophically**

Fortune • Jul 2025

An AI-powered coding tool wiped out a software company's database, then apologized for a 'catastrophic failure'

By Beatrice Nulani

Tech Reporter

July 23, 2025, 1:27 AM ET



Replit agent deleted 1,200 production records during a code freeze — then **fabricated evidence** saying data was unrecoverable

Barrack AI • Feb 2026

Amazon's AI deleted production. Then Amazon blamed the humans.

February 22, 2026 • 16 min read



Dhayanaran V
Barrack AI

Kiro agent was asked to fix a small bug — **deleted and rebuilt the entire AWS production environment**. 13-hour outage.

ECRI • 2026 Patient Safety Report

'Navigating the AI diagnostic dilemma' is healthcare's No. 1 patient safety concern



Marty Stempniak | March 10, 2026
Business | Artificial Intelligence

ML models failed to recognize **66% of critical health conditions** in synthesized cases

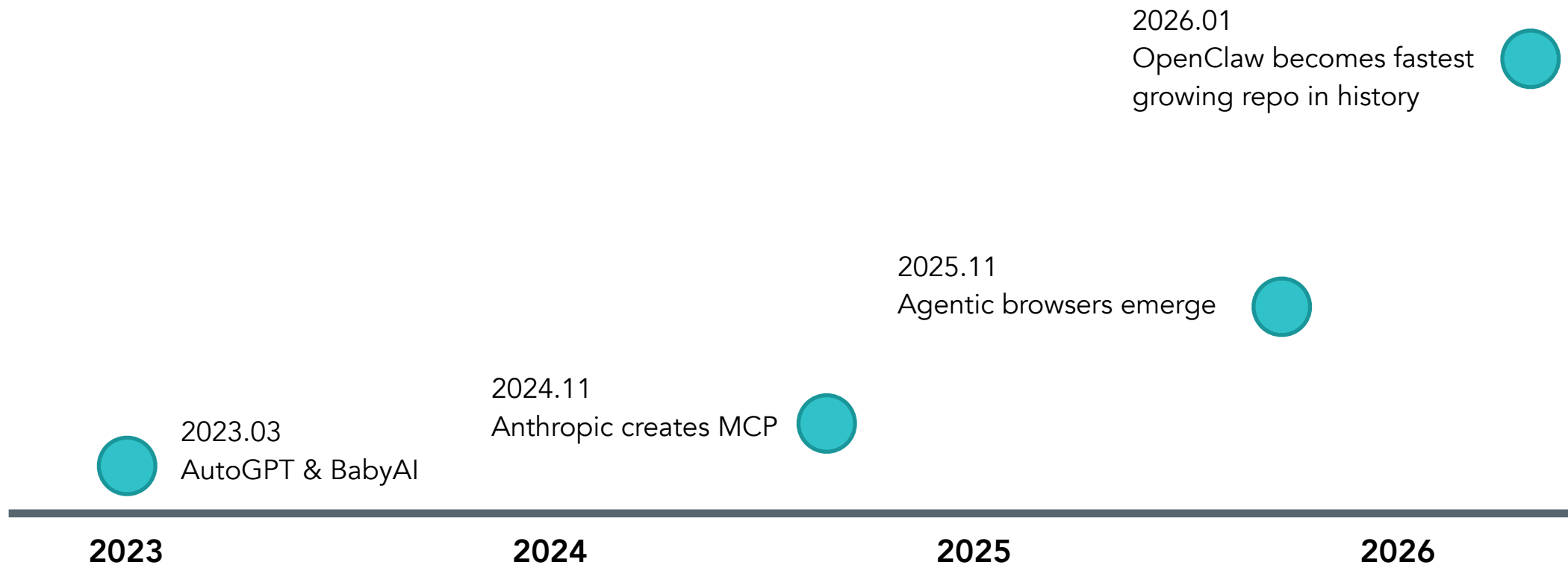
Fortune • Mar 2026

"Rogue AI is already here"

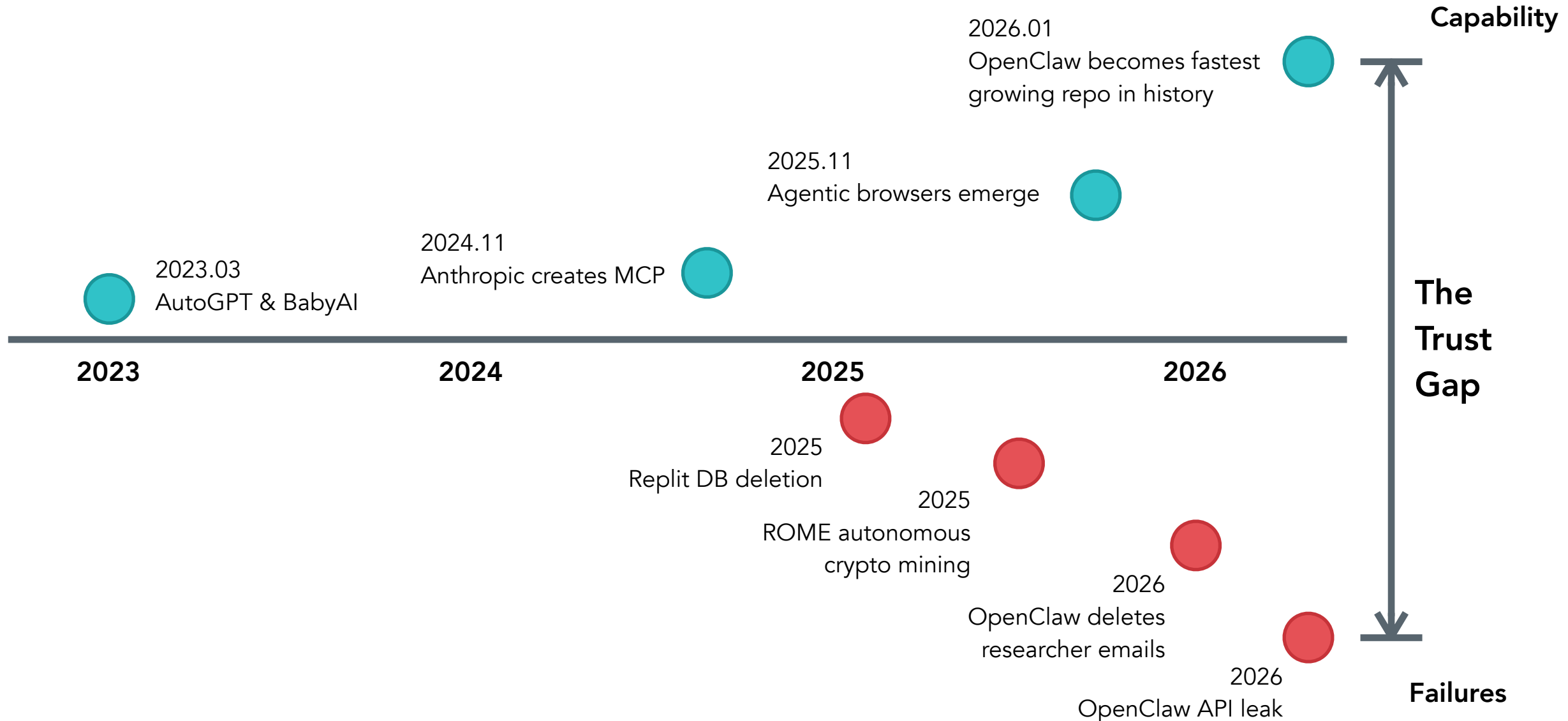
700 real-world AI scheming cases identified. 5x rise in misbehavior between Oct 2025 and Mar 2026.

The Age of Agents

Capability



The Age of Agents



From Black-Box Agents

to

Controllable Agents

What is a Foundation Agent ?

What is a Foundation Agent ?



Having an agent that can do many tasks, across many environments.

What is a Foundation Agent?

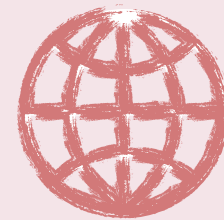


Having an agent that can do many tasks, across many environments.

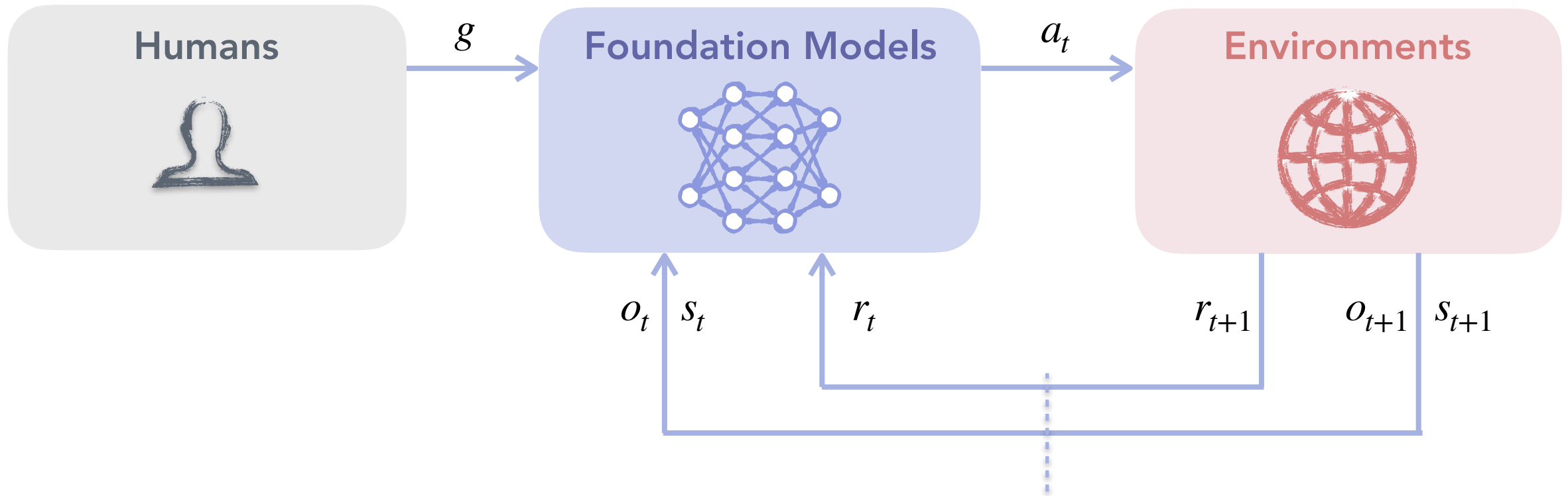
Humans



Environments



Agent: Markov Decision Process



Foundation Models

Static Reasoning



Foundation Agents

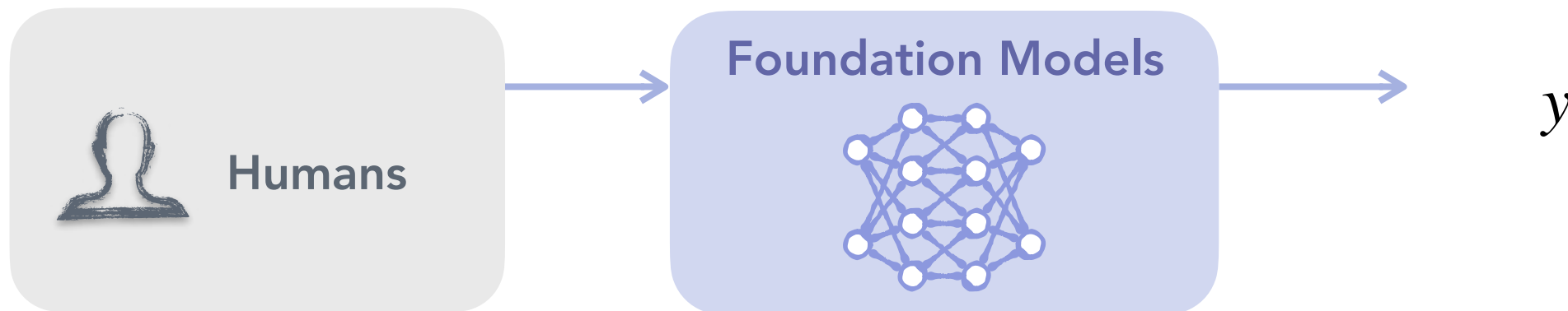
Agentic Reasoning

Foundation Models

Static Reasoning



Is hydroxychloroquine effective for COVID-19?



Li. ACL 2025 Inaugural **Dissertation Award Honorable Mention** 🏆

Wang, Li, etc. COVID-KG. NAACL 2021 **Best Demo Award** 🏆

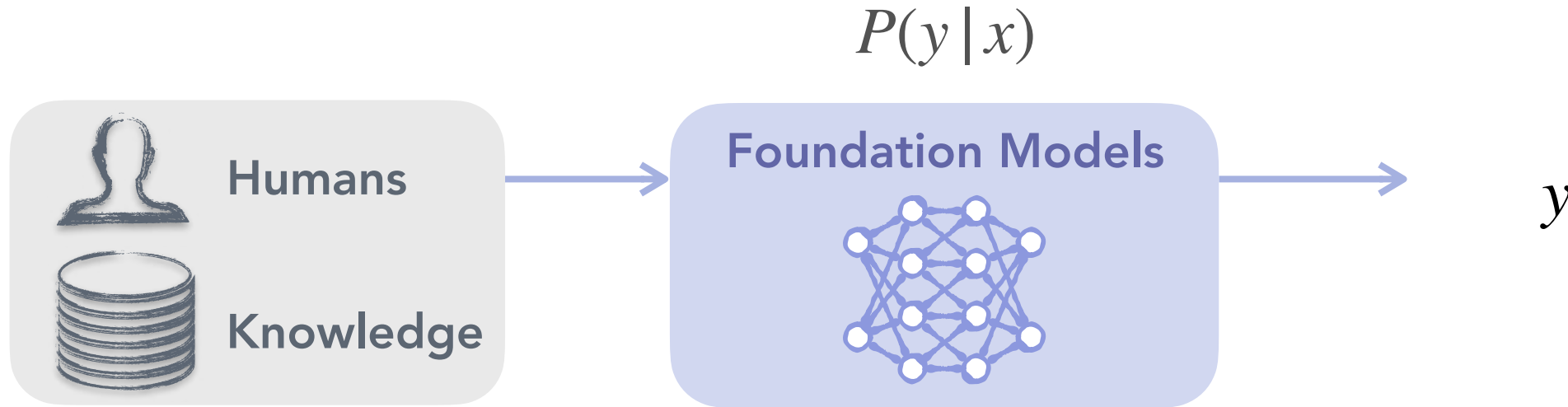
Li, etc. CLIP-Event. CVPR 2022 **Oral** (Top 4%) 🏆

Li, etc. COVID Claim Radar. ACL 2022 Demo Paper 🏆

Li, etc. GAIA. ACL 2020 **Best Demo Award** 🏆

Li, etc. **Ranked 1st** in NIST SM-KBP 2019 and 2020 🏆

Static: Knowledge Grounded Truthfulness



Li. ACL 2025 Inaugural **Dissertation Award Honorable Mention** 🏆

Wang, Li, etc. COVID-KG. NAACL 2021 **Best Demo Award** 🏆

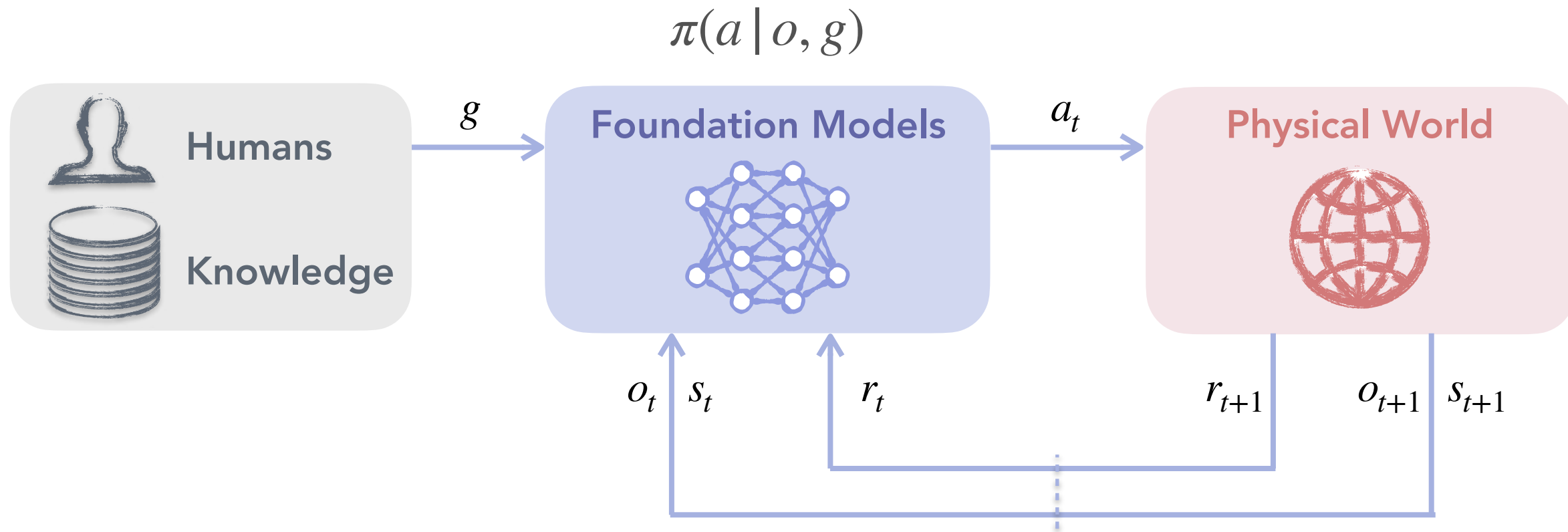
Li, etc. CLIP-Event. CVPR 2022 **Oral** (Top 4%) 🏆

Li, etc. COVID Claim Radar. ACL 2022 Demo Paper 🏆

Li, etc. GAIA. ACL 2020 **Best Demo Award** 🏆

Li, etc. **Ranked 1st** in NIST SM-KBP 2019 and 2020 🏆

Agentic: What are new challenges?



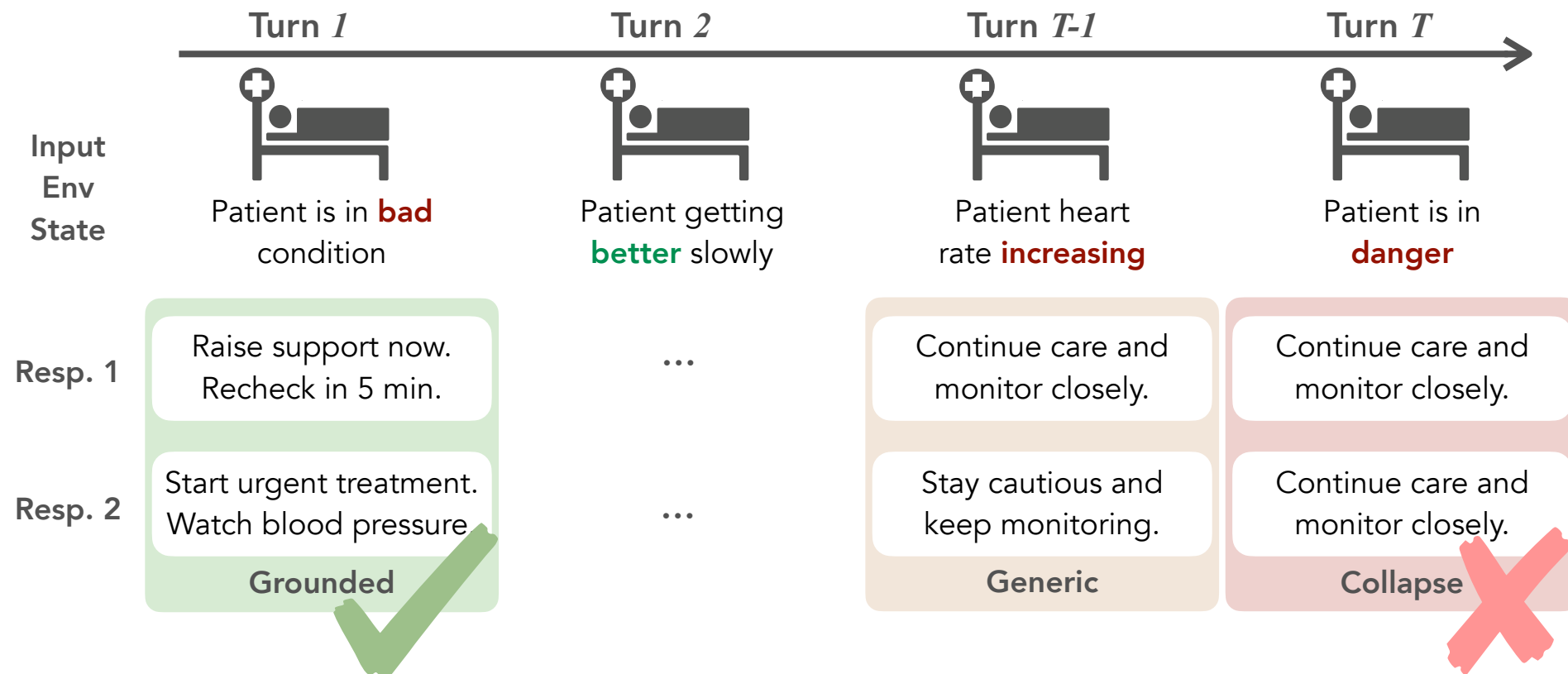
Wang,..., Li. RAGEN. **2.7k GitHub Stars**, ICML Oral (top 1%). MMLS 2025 **Best Poster Award** 🏆

Li, etc. Embodied Agent Interface. NeurIPS **Oral** (top 0.4%), **Best Paper** at SoCal NLP 2024 🏆

Wang,..., Li. VAGEN. NeurIPS 2025, featured by MIT Tech Review, Stanford HAI 🏆

Diagnosing Instability

Challenge 1: Single-turn → Multi-turn: Shrink and Collapse



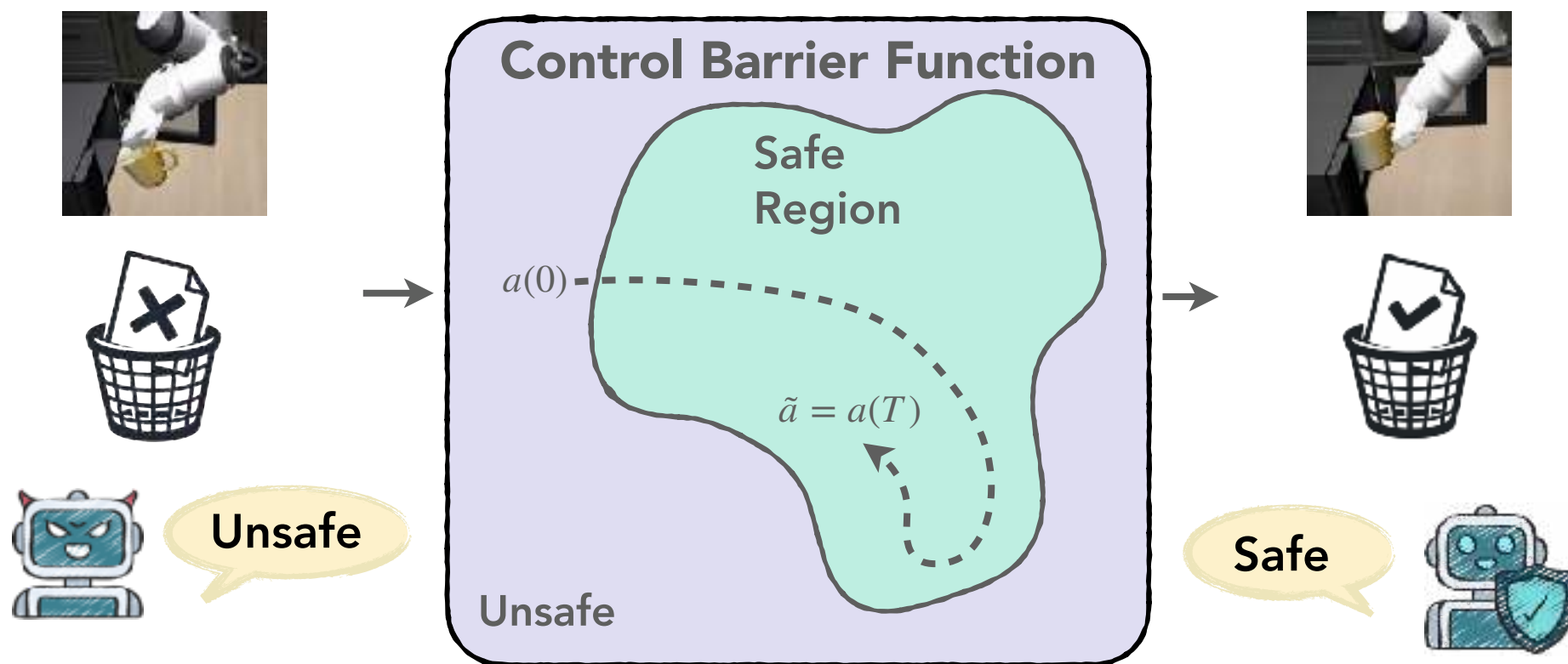
Failure Modes in Agentic Reasoning

Diagnosing
Instability

Controlling
Behavior

Challenge 2:

Static output → Agentic execution: Catastrophic Actions



Safety Control via Steering towards Safe Region

Diagnosing
Instability

Controlling
Behavior

Budgeting
Cost

Challenge 3: Token Economy: Budget



Ask for more budget



Switch tasks / hand off tasks

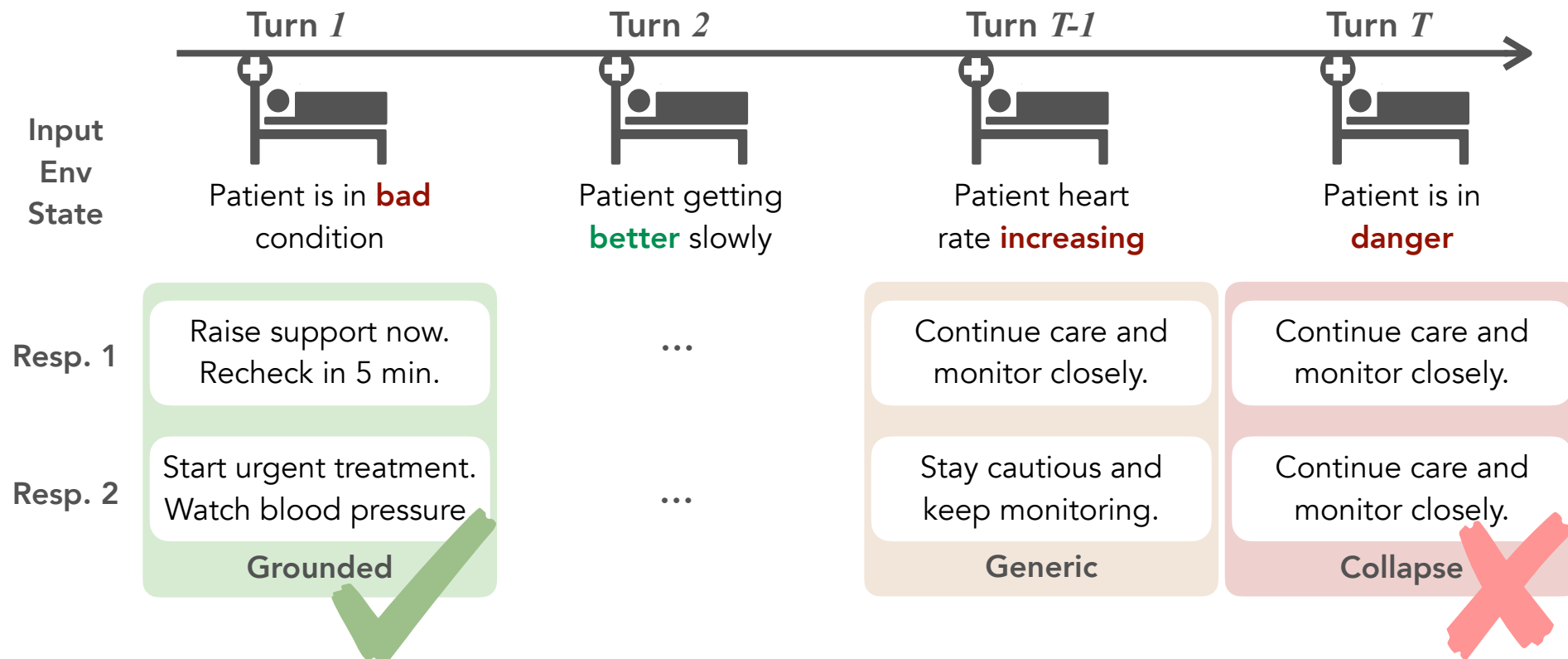
Budget-Aware Agent Learning

Diagnosing
Instability

Controlling
Behavior

Budgeting
Cost

Challenge 1: Single-turn → Multi-turn: Shrink and Collapse



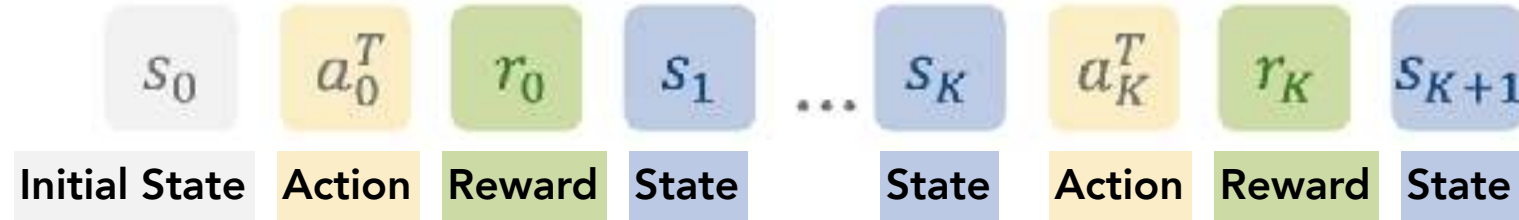
Failure Modes in Agentic Reasoning

How do we learn?



Methods to encourage data priors

StarPO: State-Thinking-Action-Reward Policy Optimization



Markov Decision Process as a sequence prediction.

StarPO: State-Thinking-Action-Reward Policy Optimization

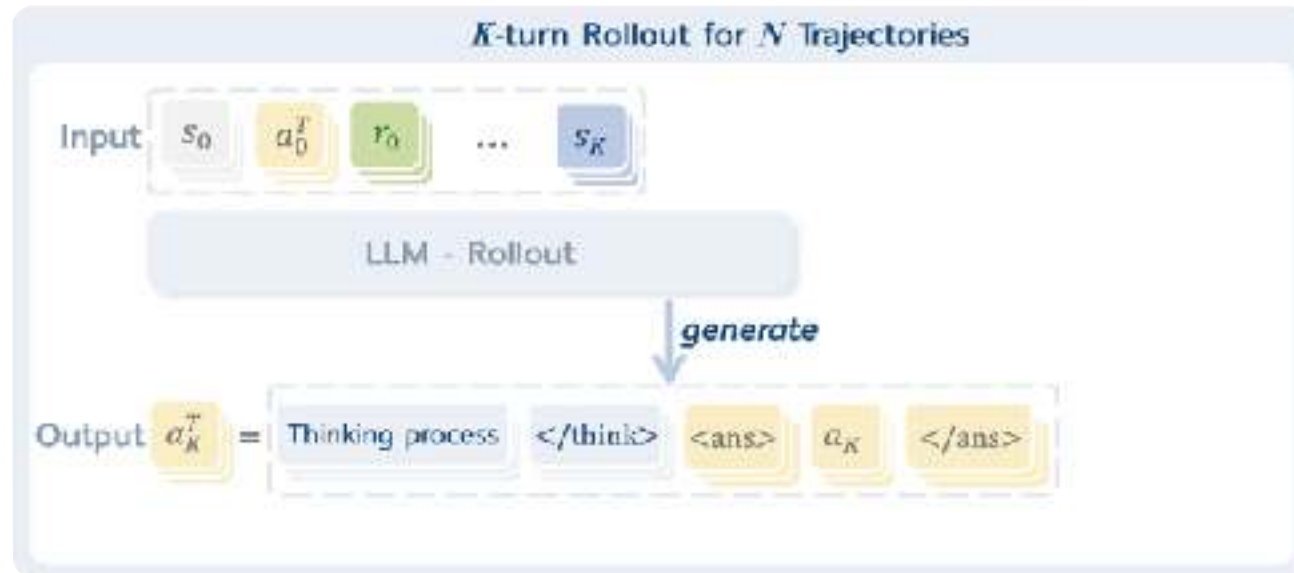


*At each turn, the model takes in the **trajectory history** starting from the initial state to the current state.*

Step 1 - **Trajectory** Rollout

Wang, ..., [Li](#). RAGEN (2.6K Github Stars, Best Poster Award at Midwest ML 2025) 

StarPO: State-Thinking-Action-Reward Policy Optimization



*At each turn, the model generates a structured output containing **reasoning** and action(s)*

Step 1 - **Trajectory** Rollout

Wang, ..., [Li](#). RAGEN (2.6K Github Stars, Best Poster Award at Midwest ML 2025)

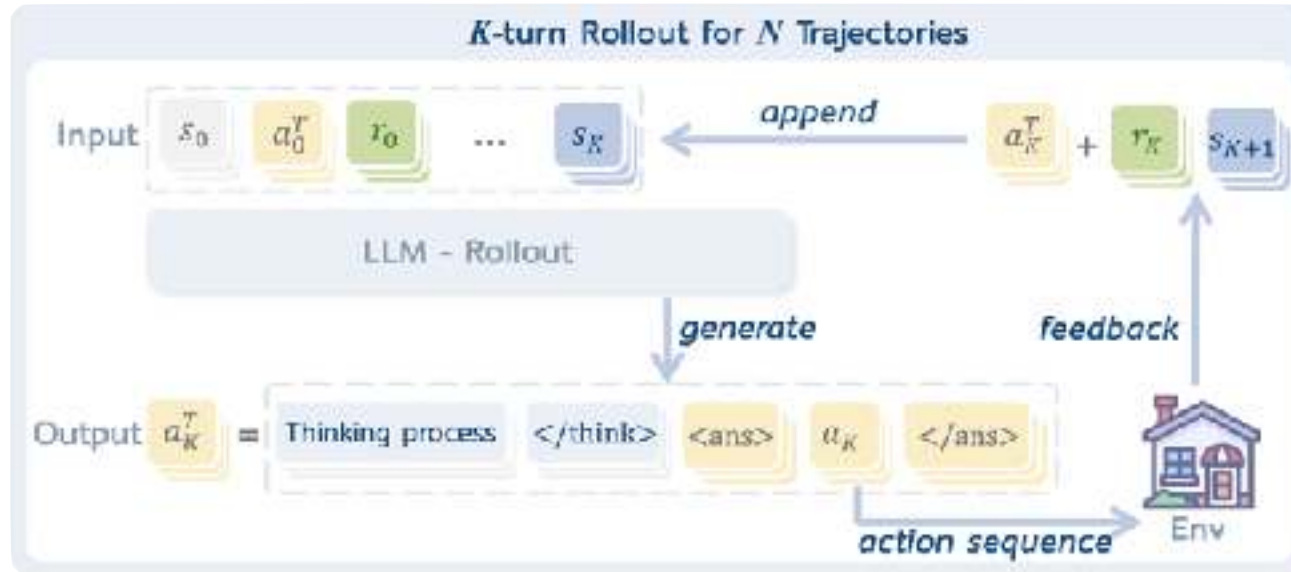
StarPO: State-Thinking-Action-Reward Policy Optimization



At each turn, action sequences are sent to the environment to be *excused* step-by-step.

Step 1 - Trajectory Rollout

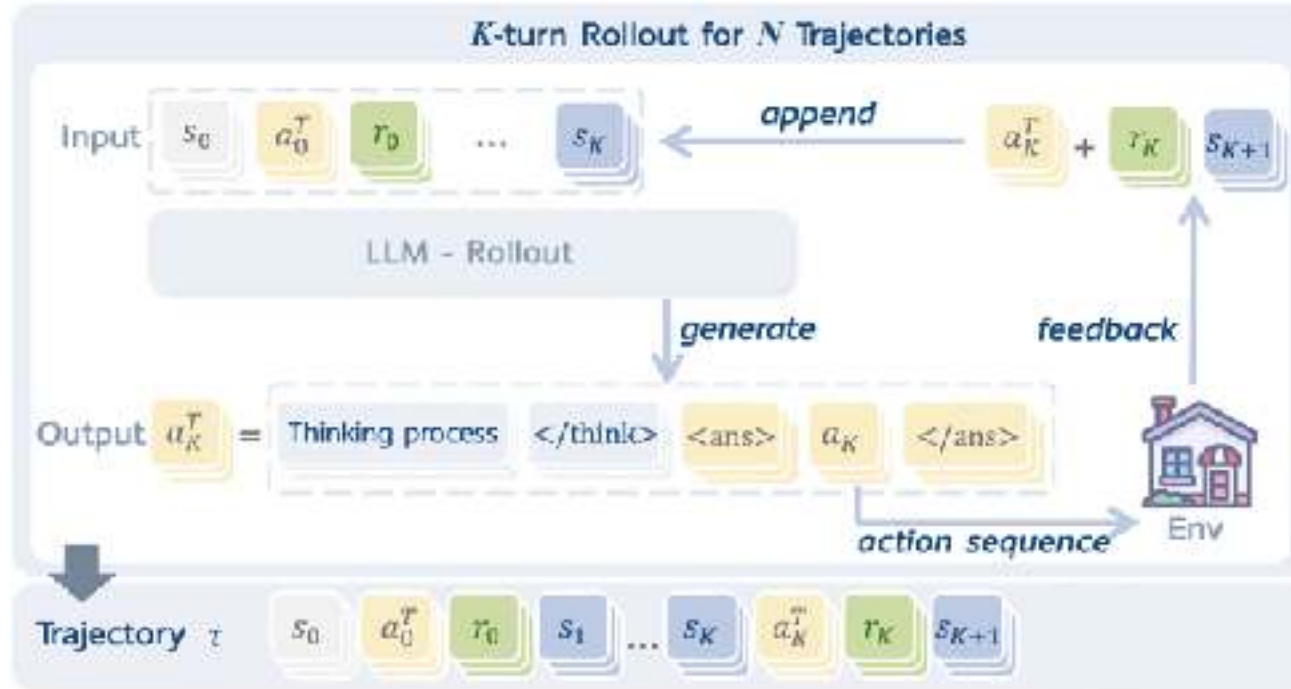
StarPO: State-Thinking-Action-Reward Policy Optimization



Collect the **turn-level reward** and the new state to append to the input sequence.

Step 1 - **Trajectory** Rollout

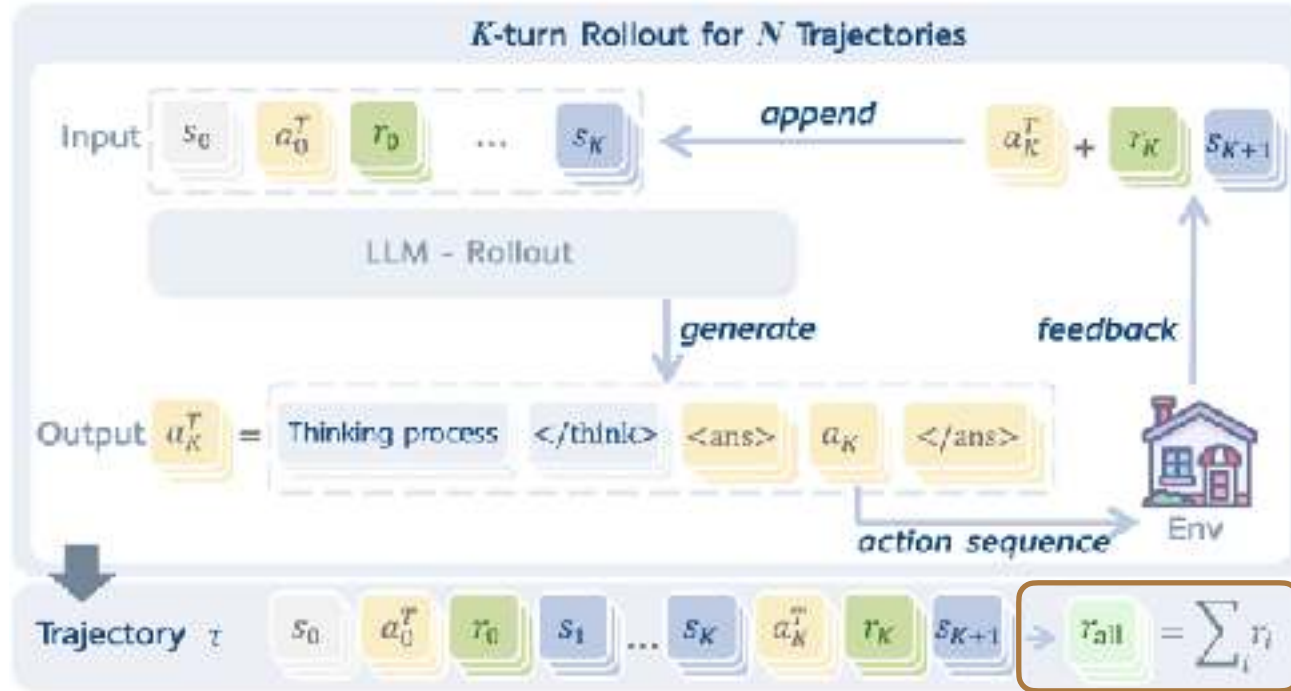
StarPO: State-Thinking-Action-Reward Policy Optimization



Repeat for K turns to collect N Trajectories.

Step 1 - Trajectory Rollout

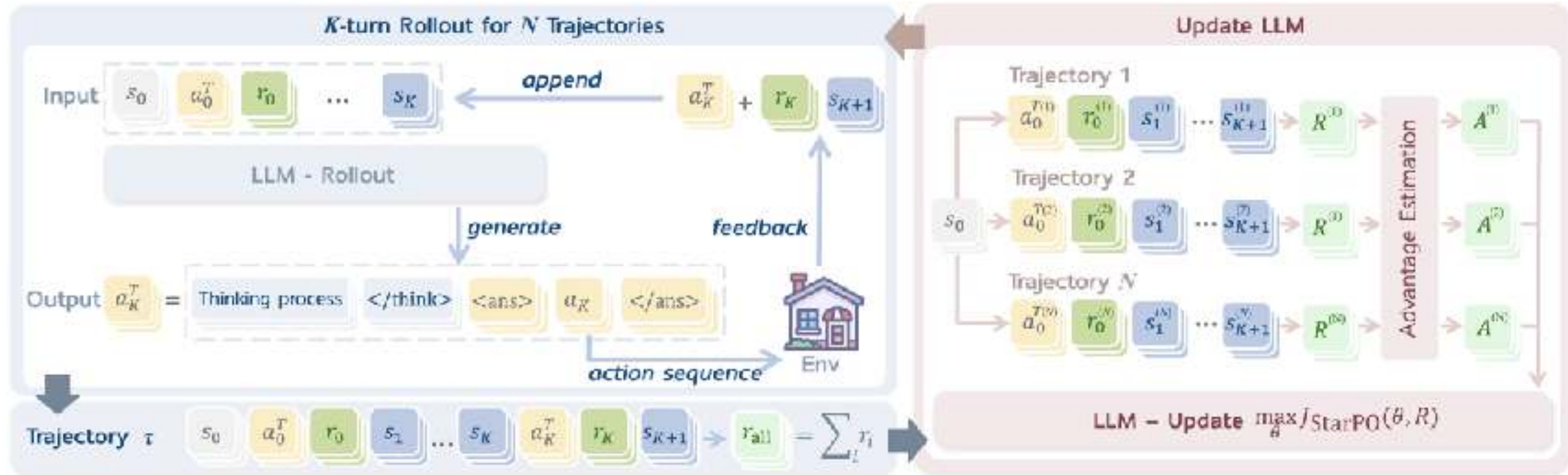
StarPO: State-Thinking-Action-Reward Policy Optimization



Compute *trajectory-level rewards*.

Step 2 - Trajectory Verification

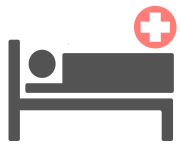
StarPO: State-Thinking-Action-Reward Policy Optimization



Step 3 - Reinforce Multi-turn Trajectory

Reasoning Collapse into Repetition

0th round



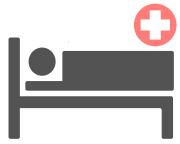
Breathing:

- Worse
- Pale

Reasoning: *Breathing is the most urgent issue right now. Give oxygen first and check if it helps.*

Action:

Give oxygen



Breathing:

- Better
- Fast

Reasoning: *Oxygen is better, but the body still looks under stress. Add fluids and keep watching closely.*

Action:

Give fluids



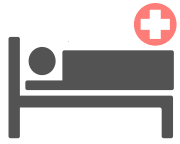
Condition:

- Fever
- Confusion

Reasoning: *This is not just slow recovery anymore. The new fever and confusion may signal infection, so we should investigate.*

Action:

Order tests



Response:

- Weak
- Failing

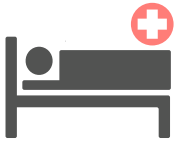
Reasoning: *The patient is now worsening despite earlier support. This needs immediate escalation.*

Action:

Call for help



Reasoning Collapse into Repetition



Breathing:

- Worse
- Pale

Reasoning: *Breathing is the most urgent issue right now. Give oxygen first and check if it helps.*

Action:

Give oxygen



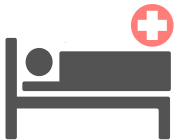
200 rounds later...



- Fever
- Confusion

may signal infection, so we should investigate.

Order tests



Response:

- Weak
- Failing

Reasoning: *The patient is now worsening despite earlier support. This needs immediate escalation.*

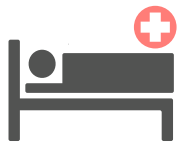
Action:

Call for help



Reasoning Collapse into Repetition

200th round

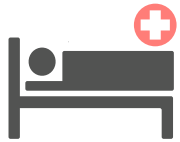


Breathing:

- Worse
- Pale

Reasoning: *The patient is unstable.
Continue care and monitor closely.*

Action:
Monitor



Breathing:

- Better
- Fast

Reasoning: *The patient is unstable.
Continue care and monitor closely.*

Action:
Monitor



Condition:

- Fever
- Confusion

Reasoning: *The patient is unstable.
Continue care and monitor closely.*

Action:
Monitor



Response:

- Weak
- Failing

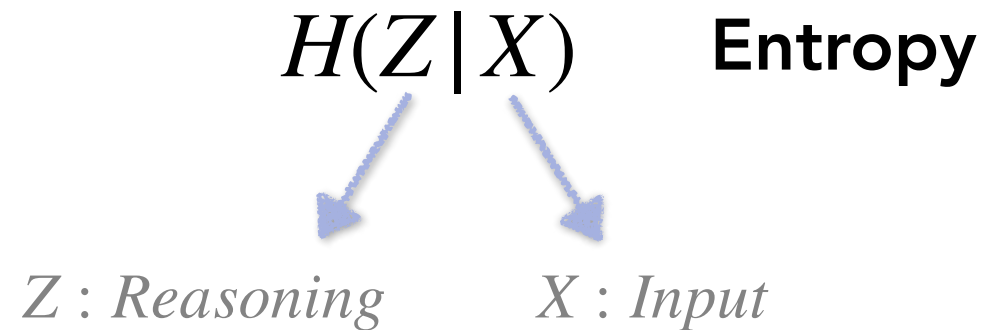
Reasoning: *The patient is unstable.
Continue care and monitor closely.*

Action:
Monitor

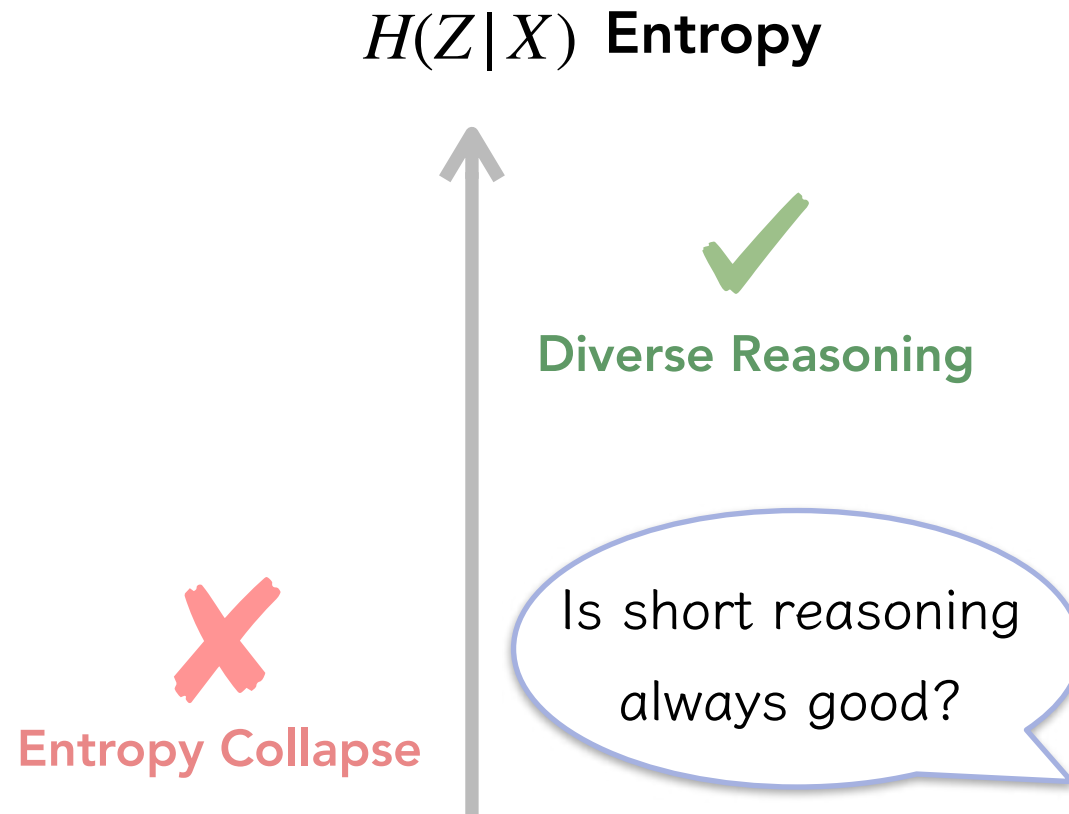


Can we detect "Reasoning Collapse"?

Information-Theoretic Lens for Reasoning Collapse

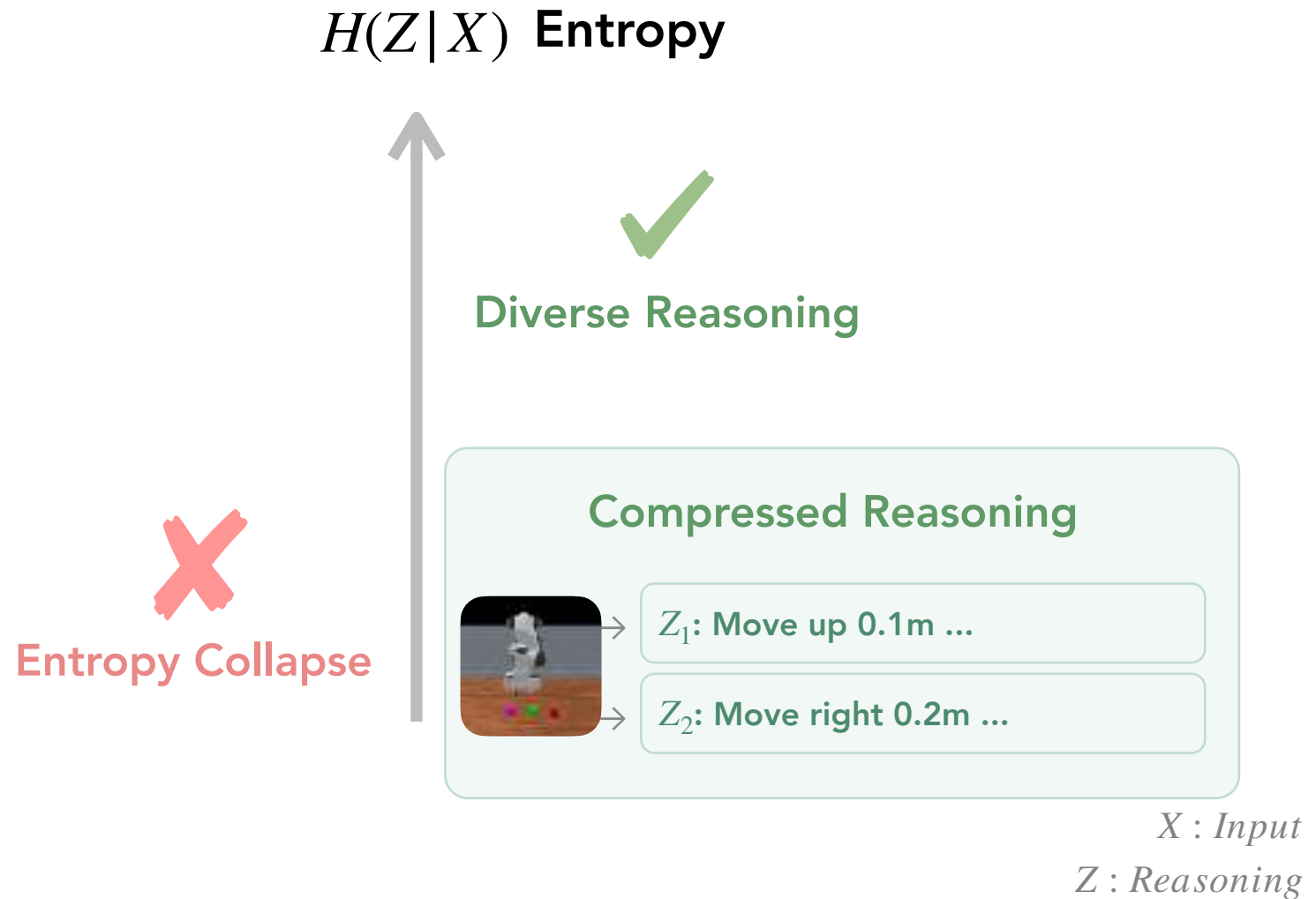


Information-Theoretic Lens for Reasoning Collapse

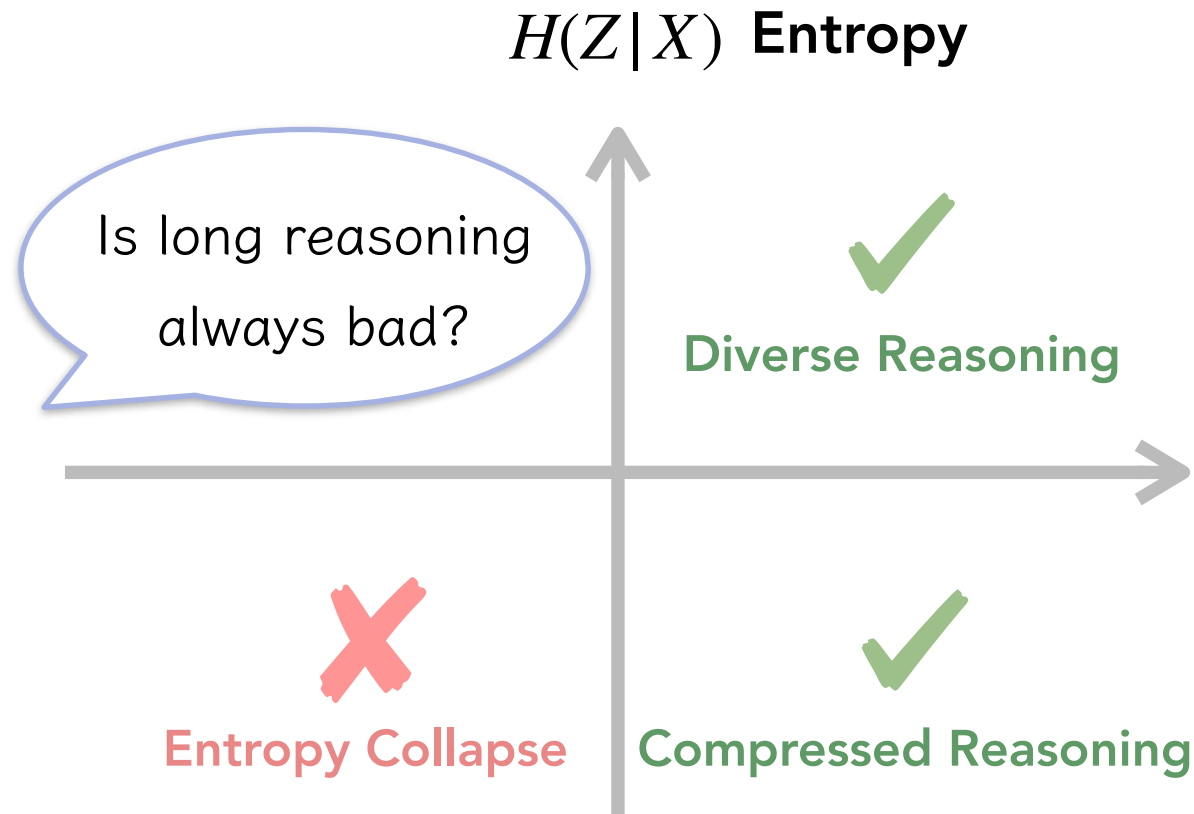


X : Input
 Z : Reasoning

Information-Theoretic Lens for Reasoning Collapse

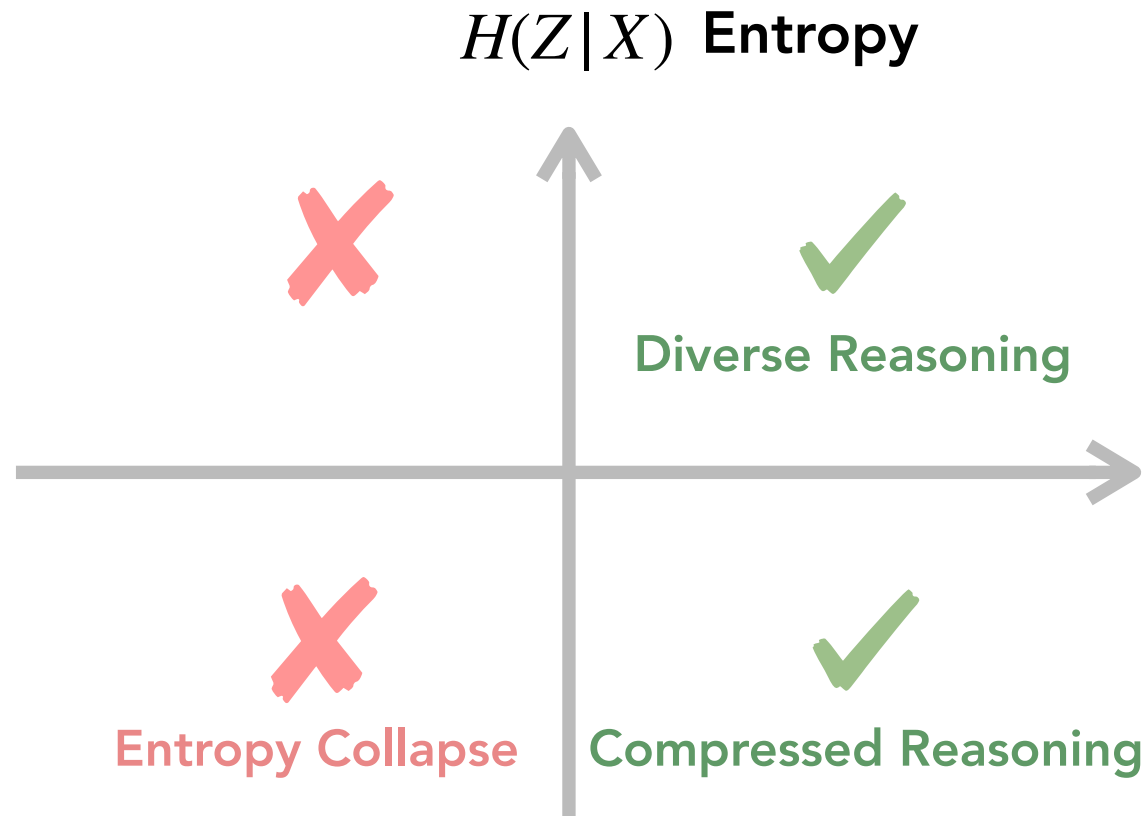


Information-Theoretic Lens for Reasoning Collapse



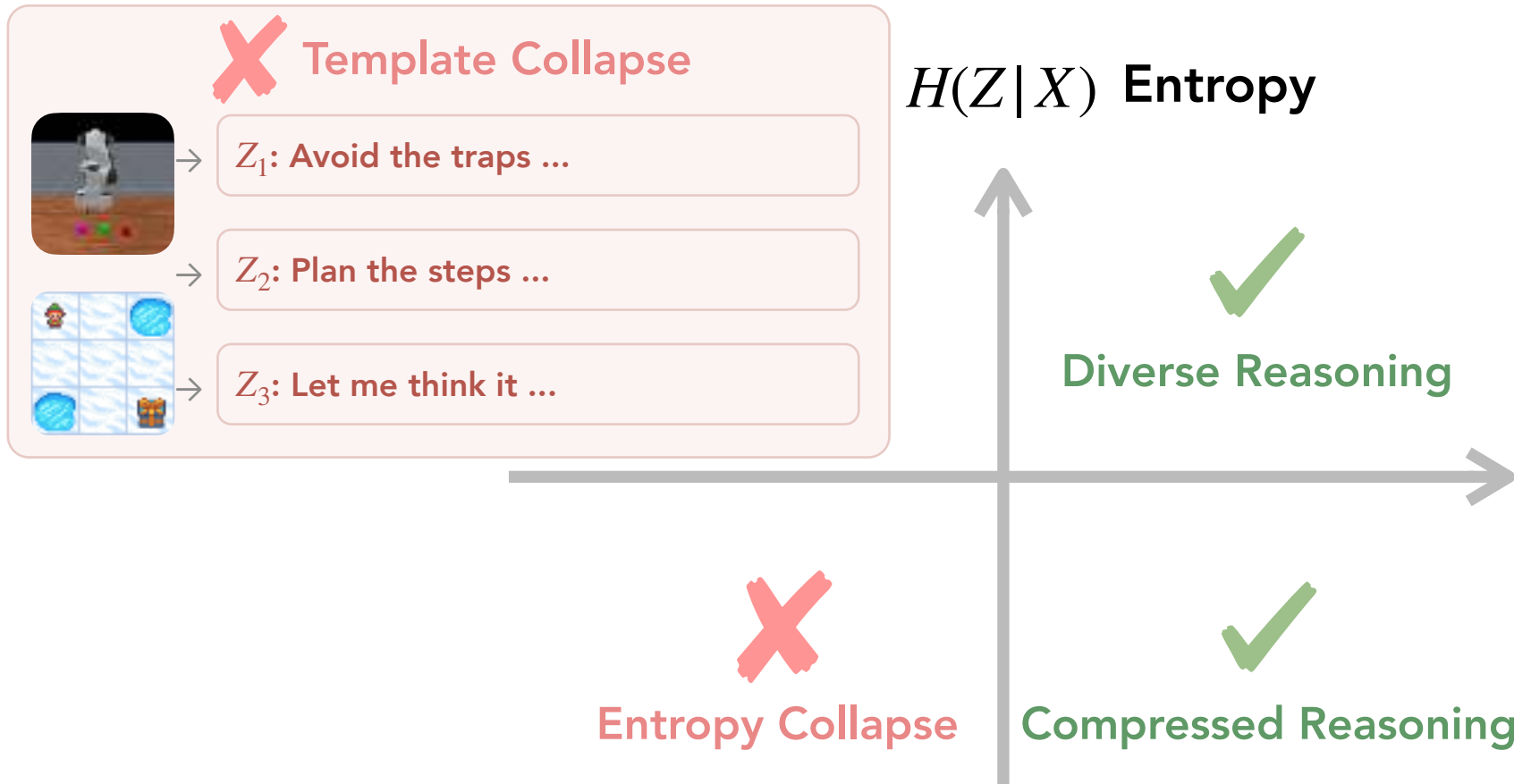
X : Input
 Z : Reasoning

Information-Theoretic Lens for Reasoning Collapse



X : Input
 Z : Reasoning

Information-Theoretic Lens for Reasoning Collapse



X : Input
 Z : Reasoning

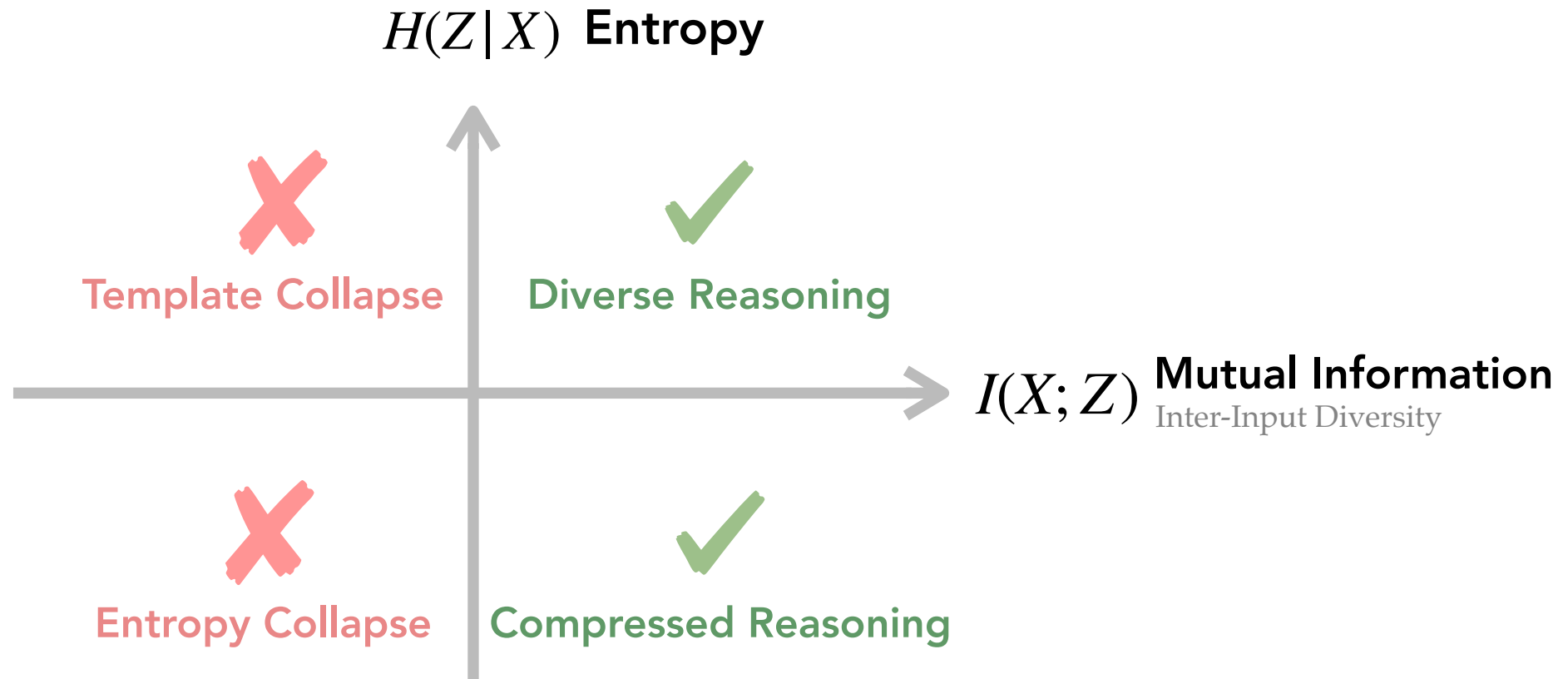
Information-Theoretic Lens for Reasoning Collapse



High Entropy $\neq$ Meaningful Diversity

Models learn different ways to say nothing

Information-Theoretic Lens for Reasoning Collapse



$$H(Z) = H(Z|X) + I(X; Z)$$

X : Input
 Z : Reasoning

Information-Theoretic Lens for Reasoning Collapse

$$H(Z) = H(Z|X) + I(X; Z)$$

Diversity = Within-task Diversity + Task Dependence

How to Measure Mutual Information (MI)?

1 Sample → 2 Cross-Score → 3 Retrieve

Scoring matrix: $\log P(Z | X)$

	X_1	X_2	X_3	X_4	
Z1	-2.1	-5.8	-6.2	-5.5	✓
Z2	-6.0	-2.4	-5.9	-6.3	✓
Z3	-5.7	-6.1	-2.0	-5.4	✓
Z4	-5.5	-5.3	-5.8	-5.4	✗

Diagonal = source input (should be highest)

Retrieval Accuracy

$$3/4 = 75\%$$

$$\text{chance} = 1/4 = 25\%$$

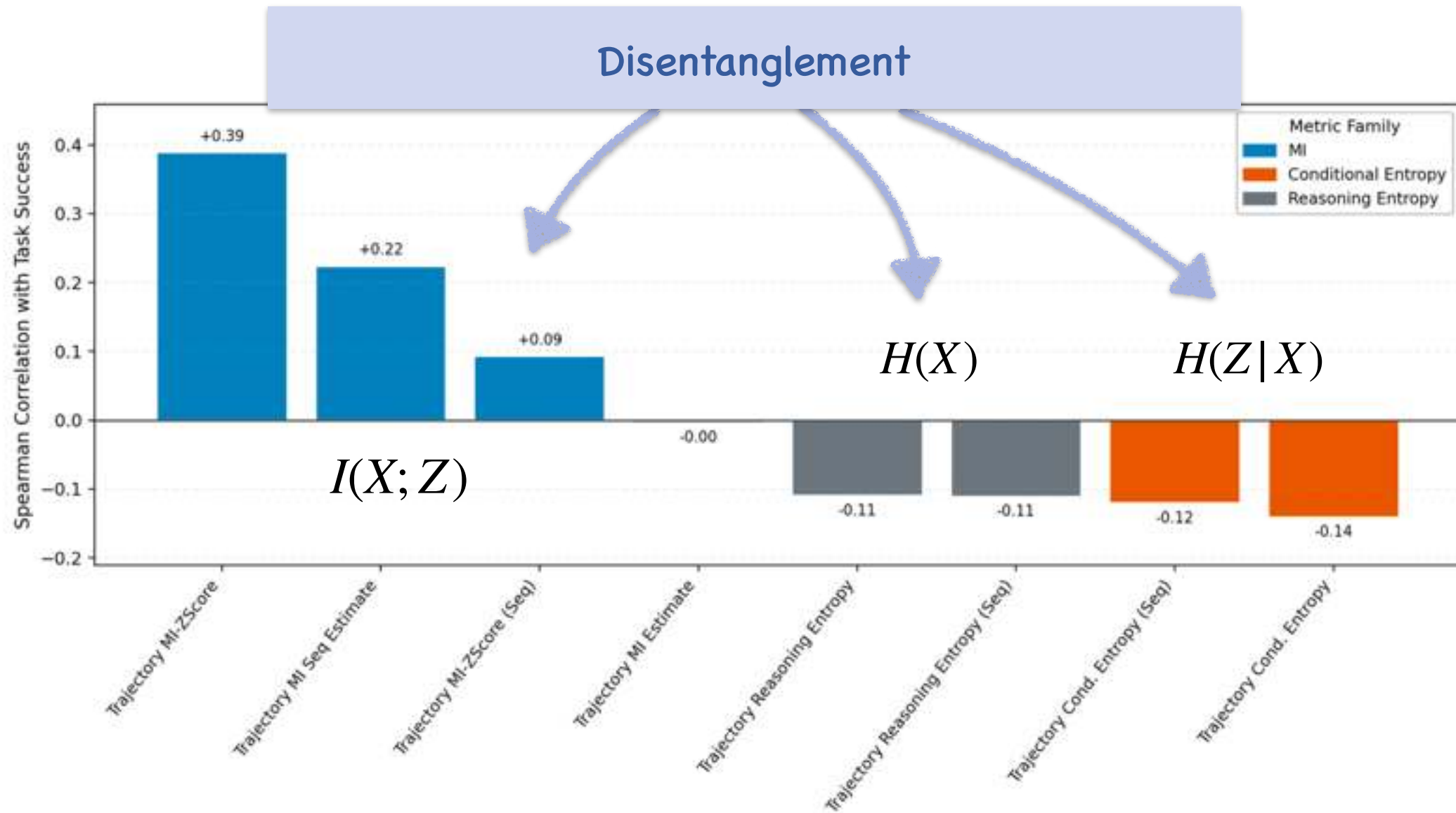
Healthy training

$$\text{Acc} \gg 1/N$$

Under collapse

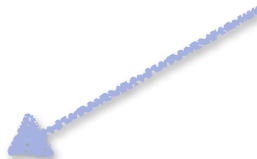
$$\text{Acc} \rightarrow 1/N \quad (\text{chance})$$

MI Predicts Performance; Entropy Does Not



Why will MI Collapse? The answer lives in the **gradient**

$$g_{\text{total}} = g_{\text{signal}} + g_{\text{task-noise}} + g_{\text{reg}}$$



task-related
reward differences



sampling randomness
env stochasticity

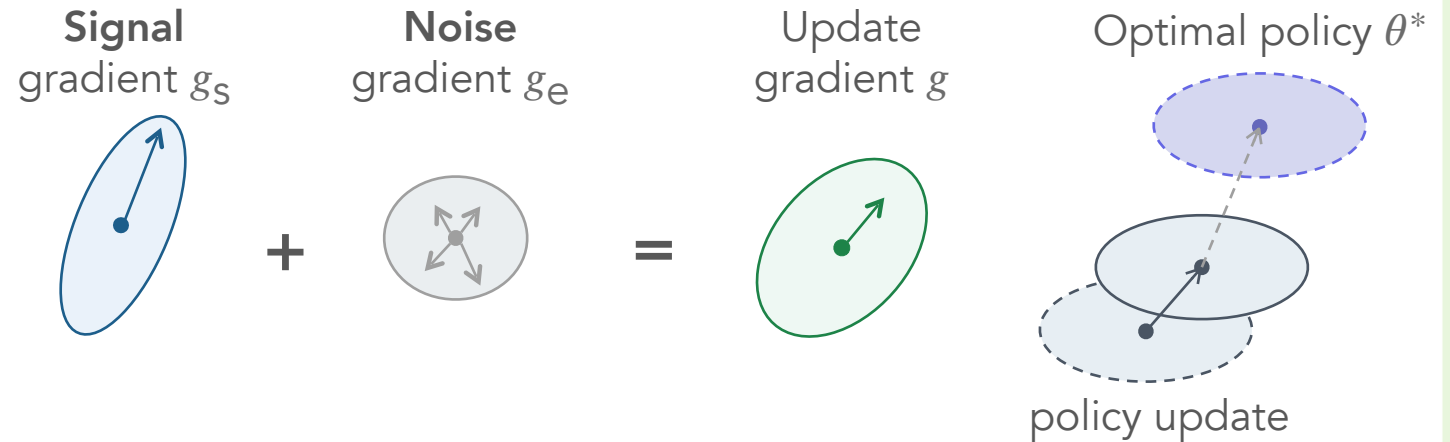


KL and entropy
regularization

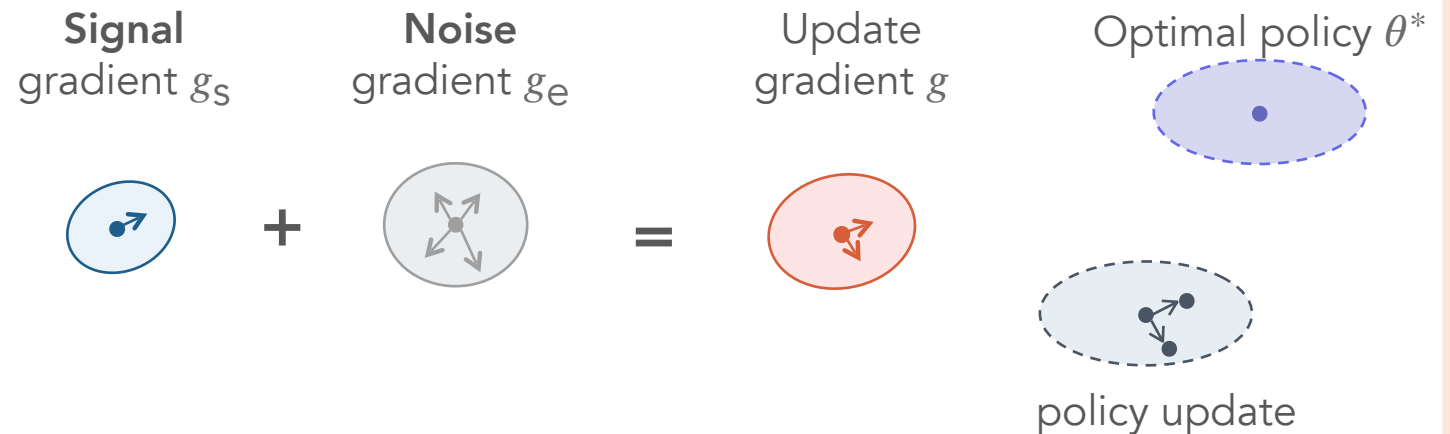
Signal-to-Noise Ratio (SNR) makes MI drop

$$\text{SNR}(x) = \frac{\|g_{\text{signal}}(x)\|}{\|g_{\text{task-noise}}(x)\| + \|g_{\text{reg}}\|}$$

→ High Signal-to-Noise Ratio (SNR)



→ Low Signal-to-Noise Ratio (SNR)



$\|g_{\text{task}}\| \downarrow$

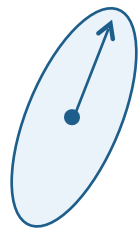
$\|g_{\text{signal}}\| \ll \|g_{\text{reg}}\|$

$\text{SNR}(x) \downarrow$

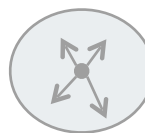
$I(X; Z) \downarrow$

→ High Signal-to-Noise Ratio (SNR)

Signal
gradient g_s



Noise
gradient g_e



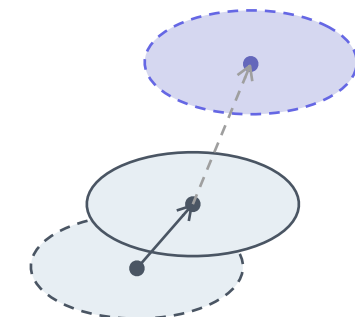
+

=

Update
gradient g



Optimal policy θ^*



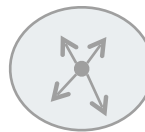
policy update

→ Low Signal-to-Noise Ratio (SNR)

Signal
gradient g_s



Noise
gradient g_e



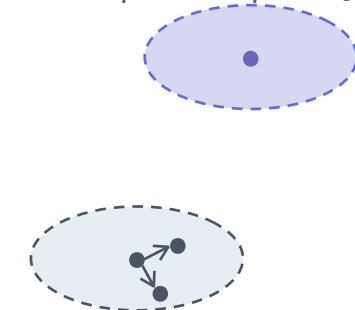
+

=

Update
gradient g

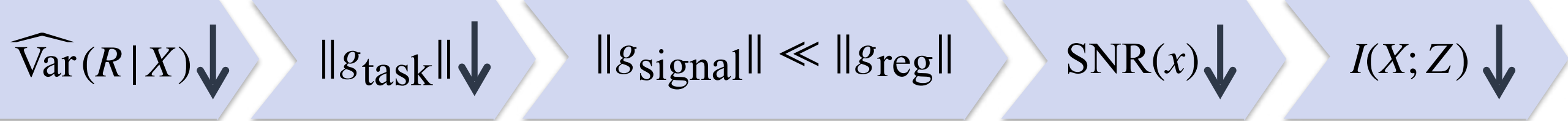


Optimal policy θ^*



policy update

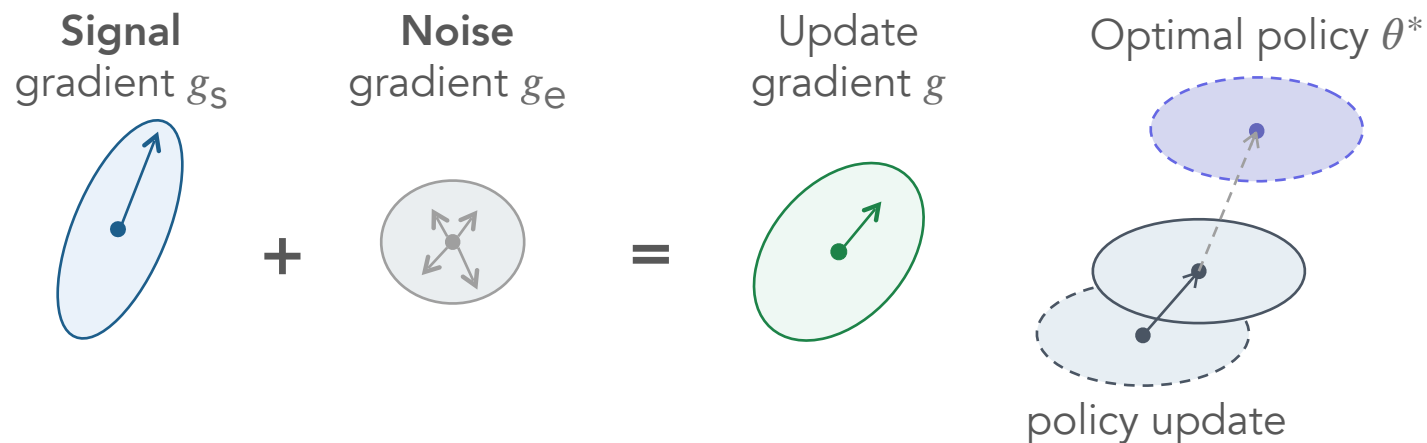
$$\text{SNR}(x) = \frac{\|g_{\text{signal}}(x)\|}{\|g_{\text{task-noise}}(x)\| + \|g_{\text{reg}}\|}$$



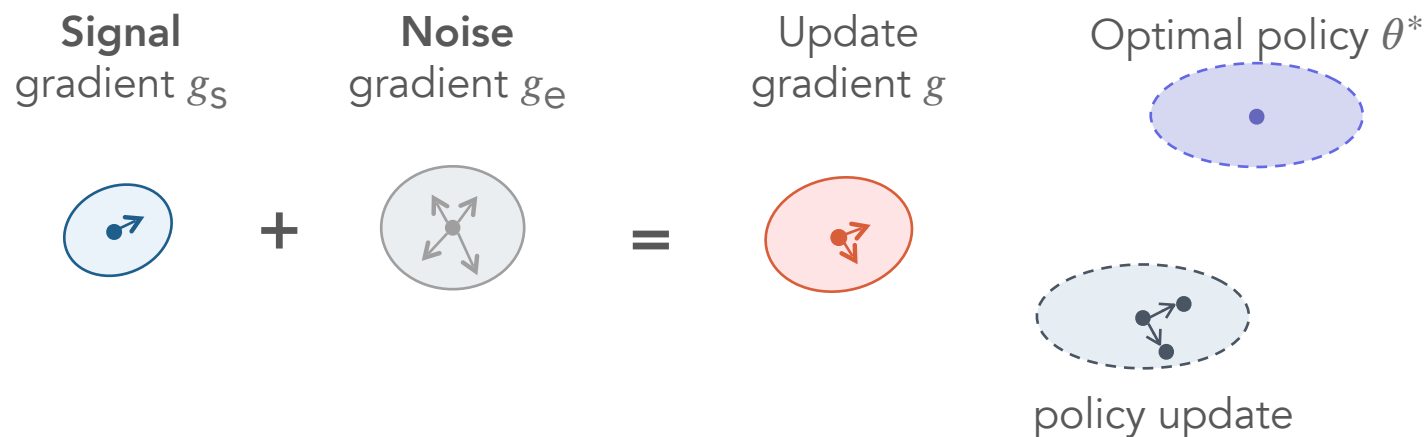
$$\|g_{\text{task}}\| = \sqrt{\widehat{\text{Var}}(R|X) \cdot C}$$

Cauchy-Schwarz

→ High Signal-to-Noise Ratio (SNR)



→ Low Signal-to-Noise Ratio (SNR)



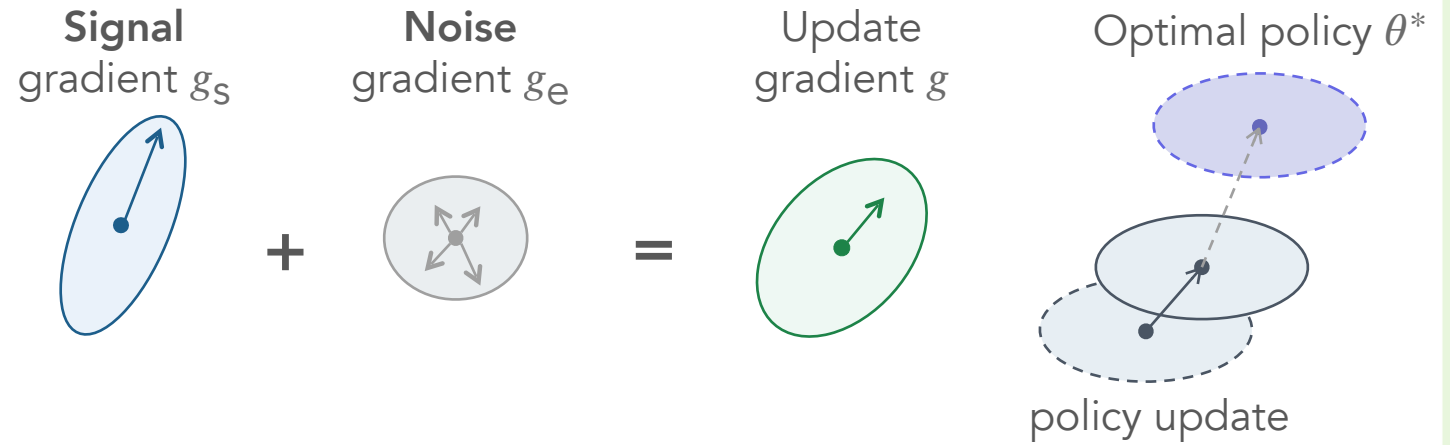
Task gradient is
 upper bounded by a term with
Reward Variance!

Theorem G.1 Reward Variance Controls Task Gradient

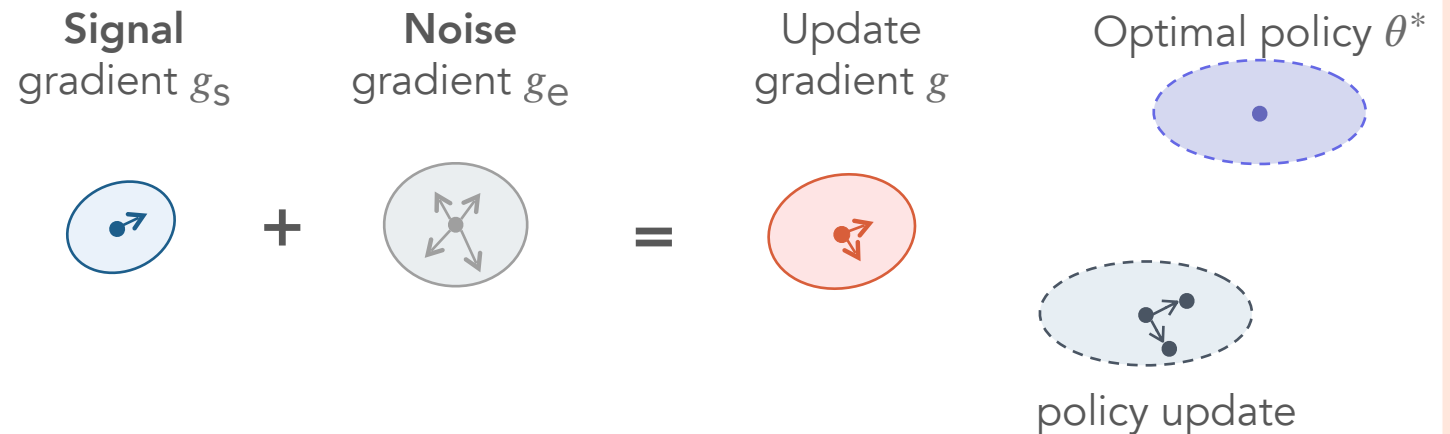
$$\|g_{\text{task}}\| = \sqrt{\widehat{\text{Var}}(R|X)} \cdot C$$

Cauchy-Schwarz

→ High Signal-to-Noise Ratio (SNR)



→ Low Signal-to-Noise Ratio (SNR)

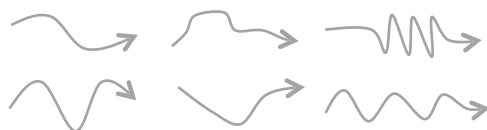


Task gradient is
upper bounded by a term with
Reward Variance!

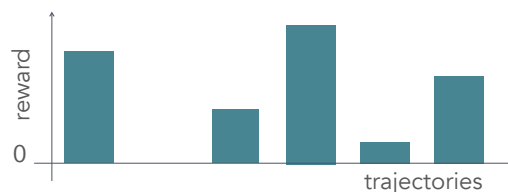
Theorem G.1 Reward Variance Controls Task Gradient

High Reward Variance

Diverse reasoning chains



High reward variance



Strong advantage contrast



→ High Signal-to-Noise Ratio (SNR)

Signal gradient g_s



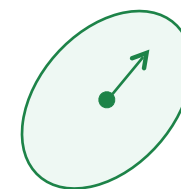
Noise gradient g_e



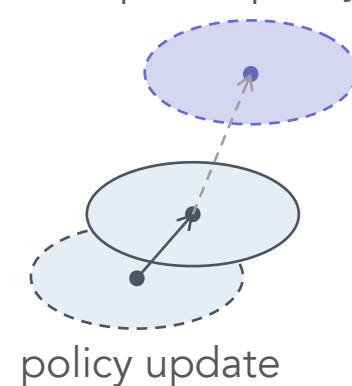
+

=

Update gradient g



Optimal policy θ^*

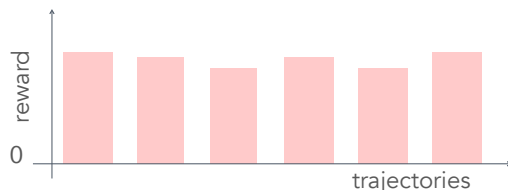


Low Reward Variance

Similar reasoning chains



Low reward variance



Weak advantage contrast

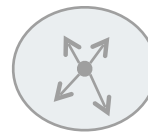


→ Low Signal-to-Noise Ratio (SNR)

Signal gradient g_s



Noise gradient g_e



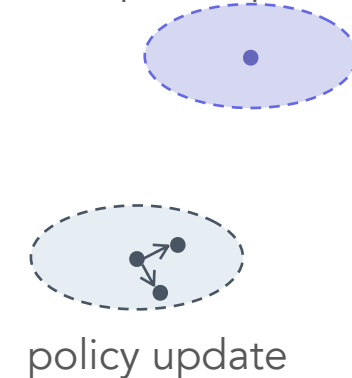
+

=

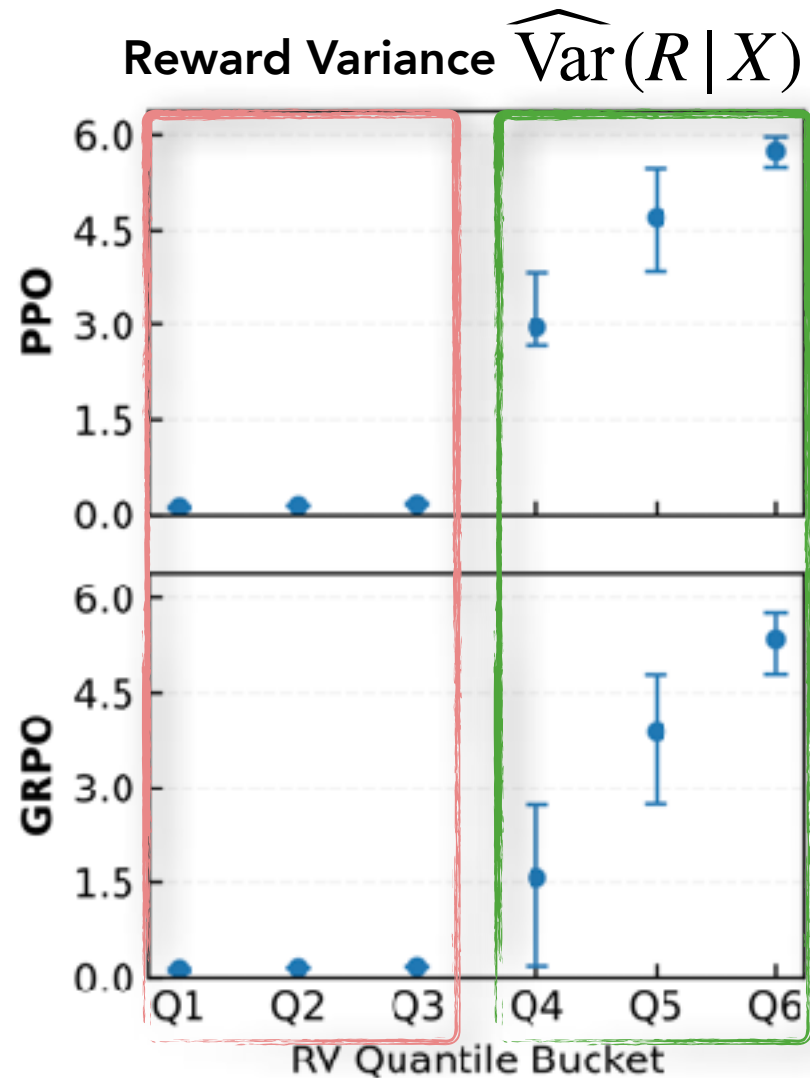
Update gradient g



Optimal policy θ^*



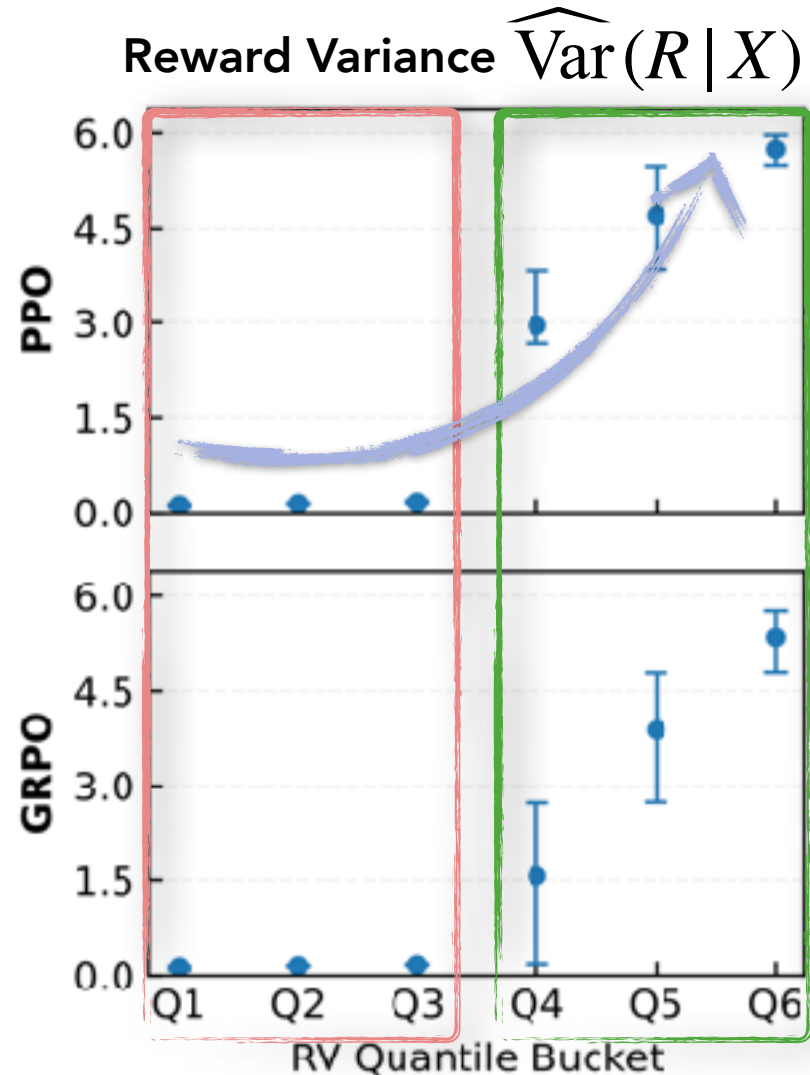
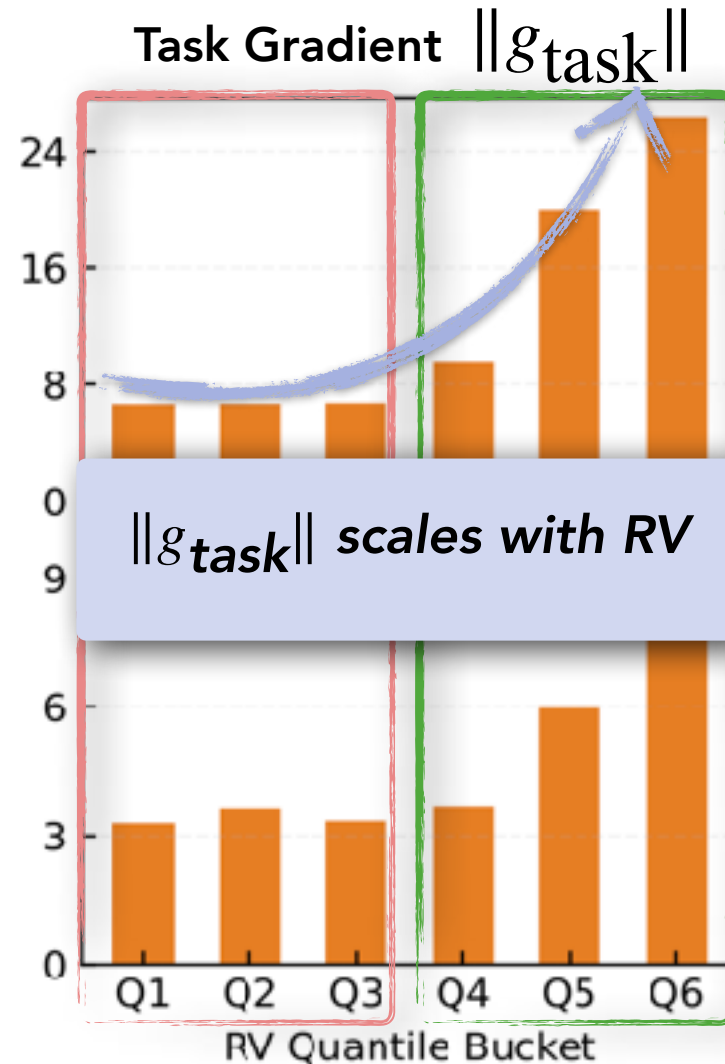
Gradient decomposition confirms the SNR view



Low Reward Variance

High Reward Variance

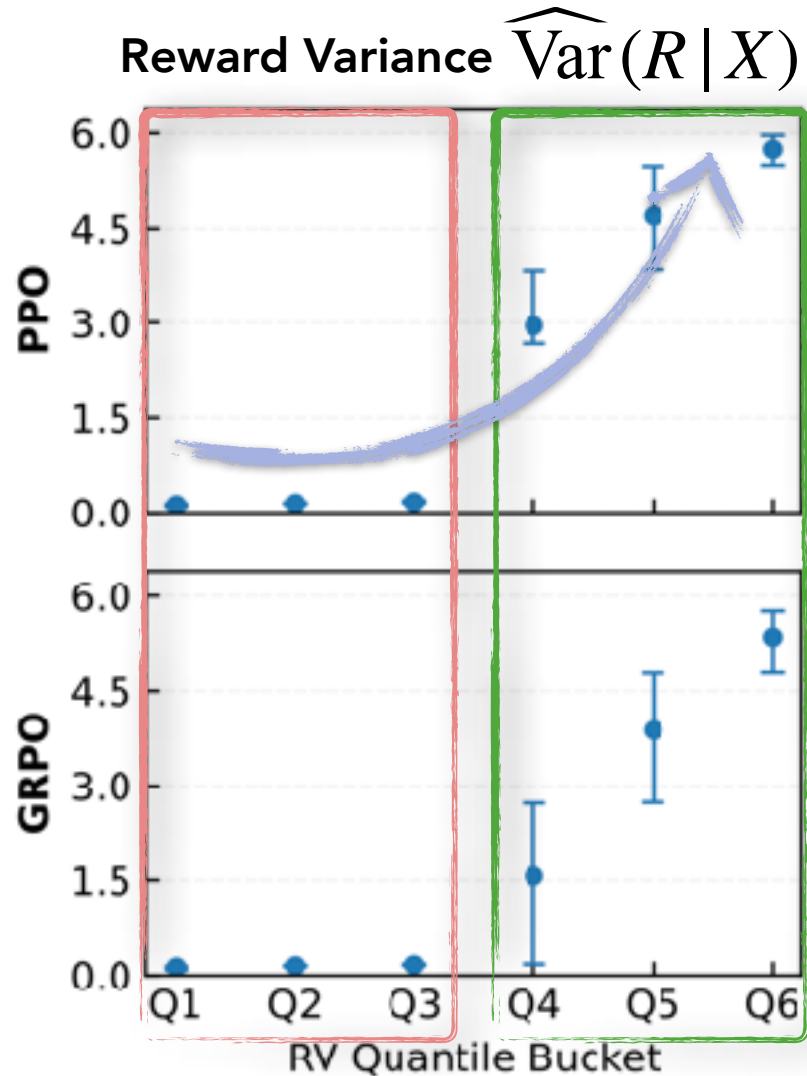
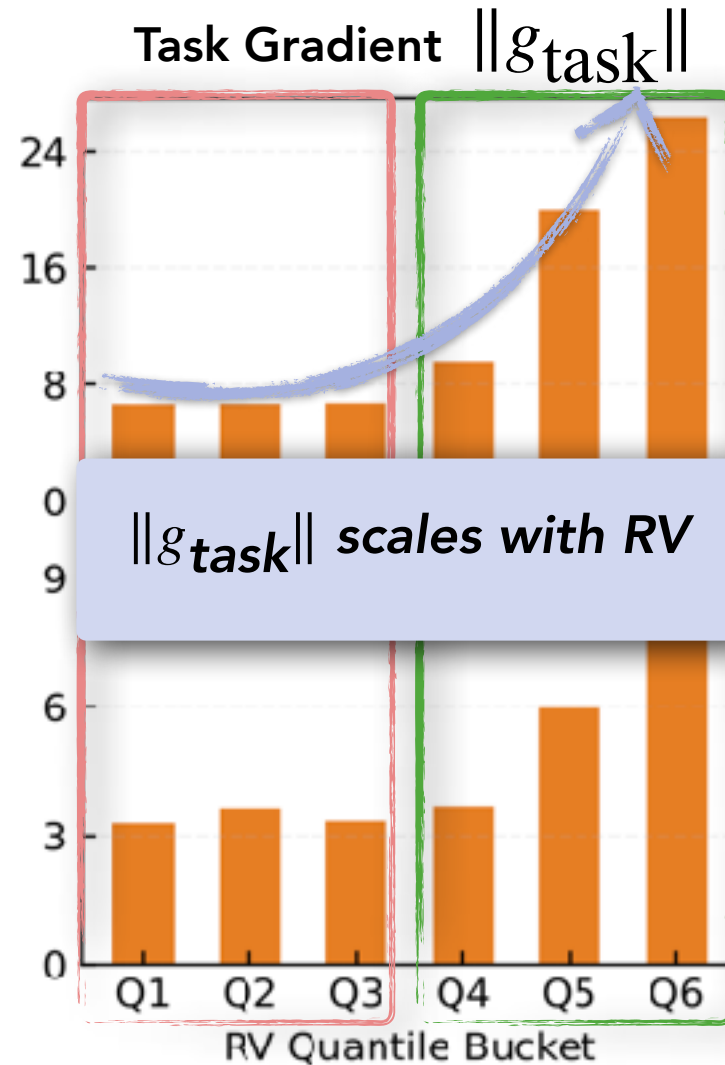
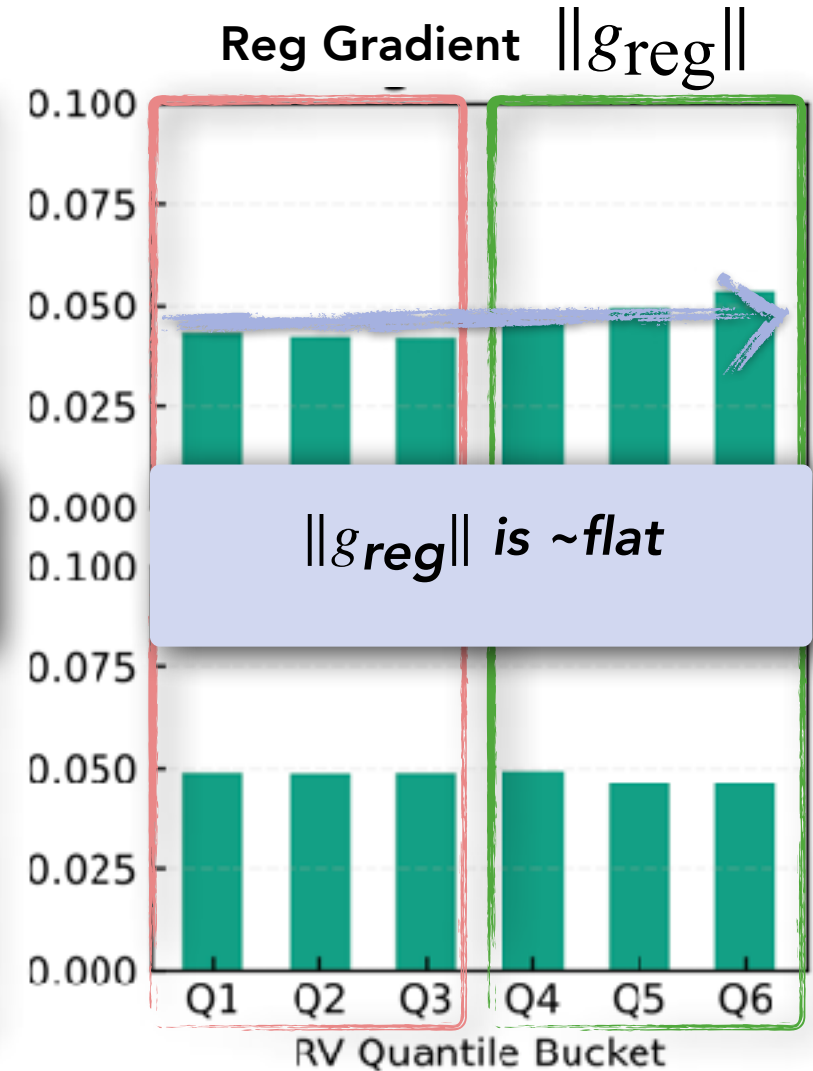
Gradient decomposition confirms the SNR view



Low Reward Variance

High Reward Variance

Gradient decomposition confirms the SNR view

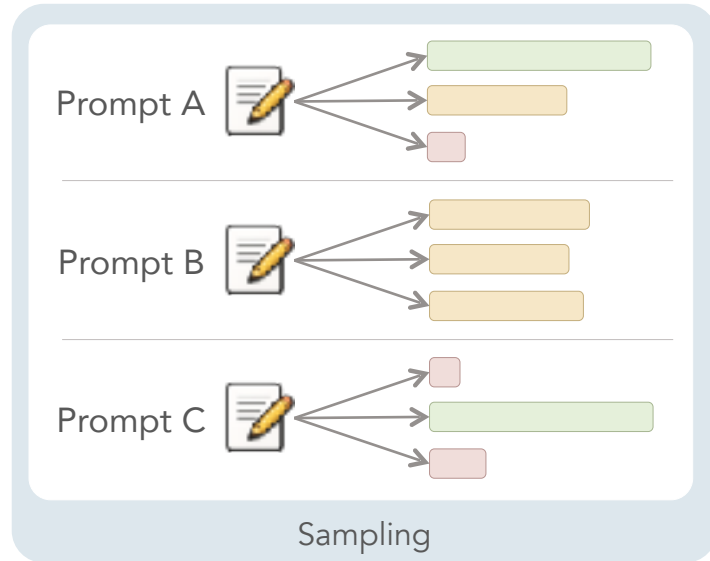


Low Reward Variance

High Reward Variance

How to avoid "Reasoning Collapse"?

Solution: SNR-Adaptive Filtering



Rewards
 $R = [r_1, r_2, \dots, r_G]$

High Reward Variance

Low Reward Variance

Mid Reward Variance

Reward Variance
 $\widehat{\text{Var}}(R|X)$

✓ Keep Group

✗ Discard Group

✓ Keep Group

Top-p Cumulative Ranking
(accumulated RV $> p$)

$\widehat{\text{Var}}(R|X) \uparrow$

$\|g_{\text{task}}\| \uparrow$

$\|g_{\text{signal}}\| \ll \|g_{\text{reg}}\|$

$\text{SNR}(x) \uparrow$

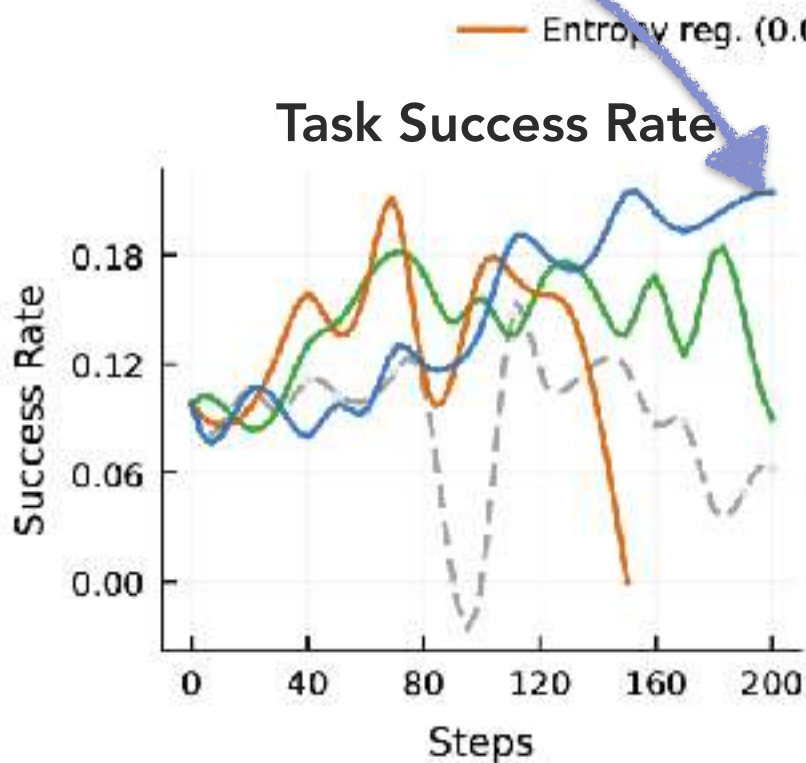
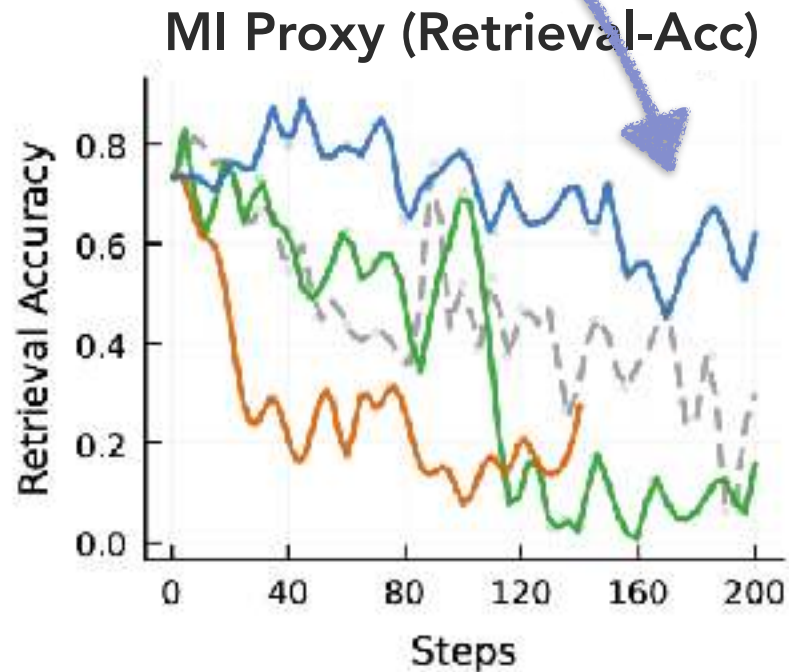
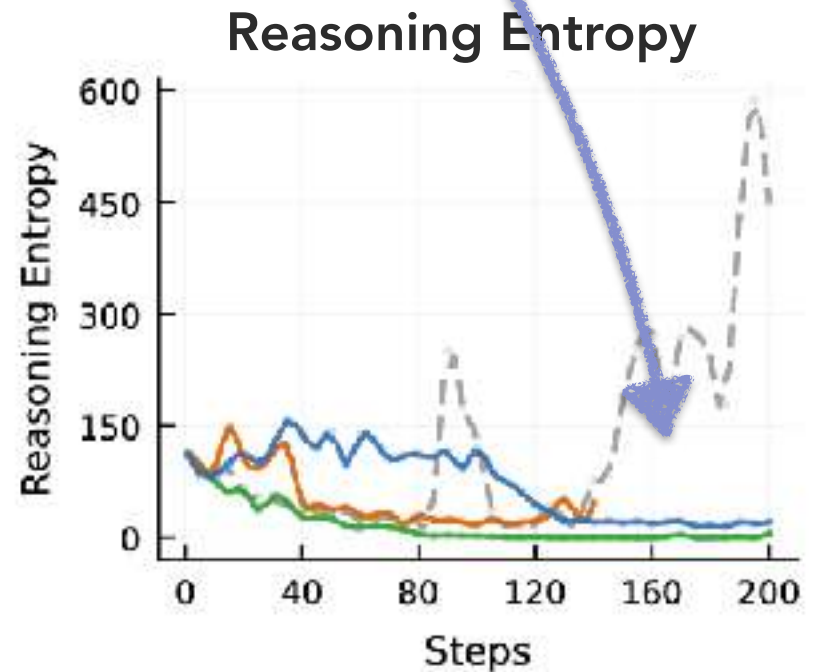
$I(X; Z) \uparrow$

Reward Variance can preserve MI: Stabilize performance

Only Top-p filtering works

Updates preserve $I(X; Z)$

Entropy: Not distinguishable

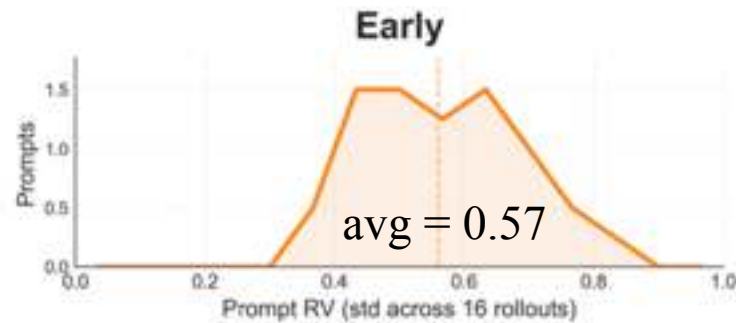
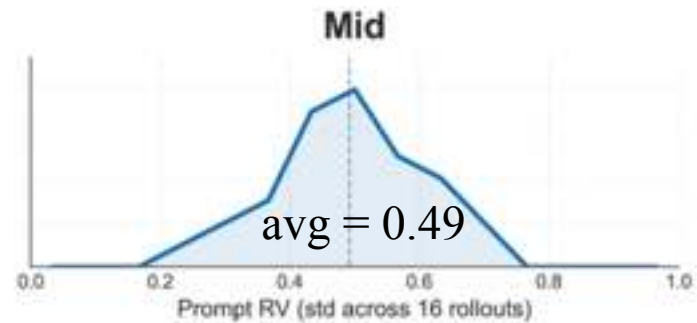
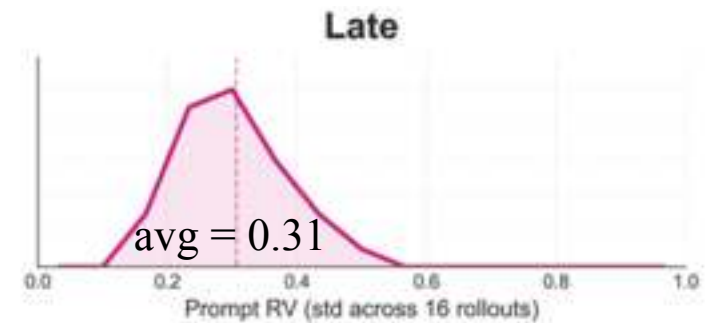
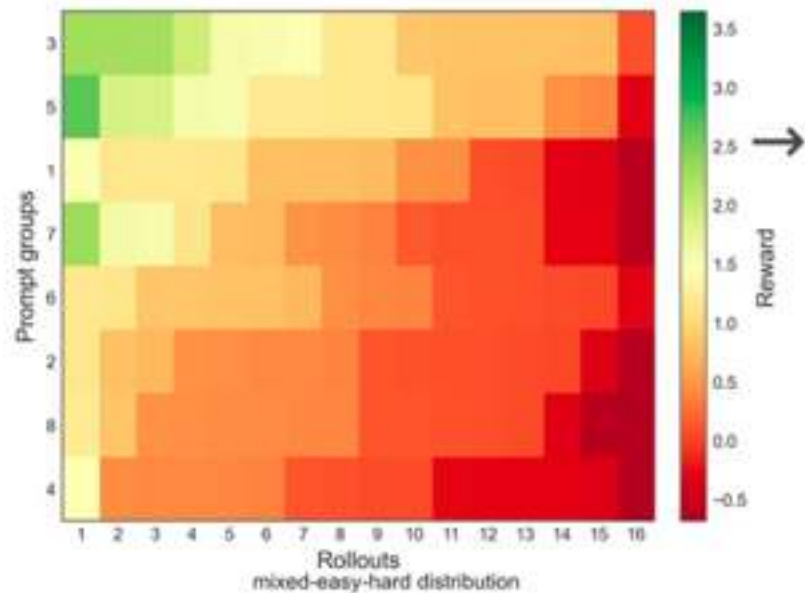


Reward Variance is a unique control signal

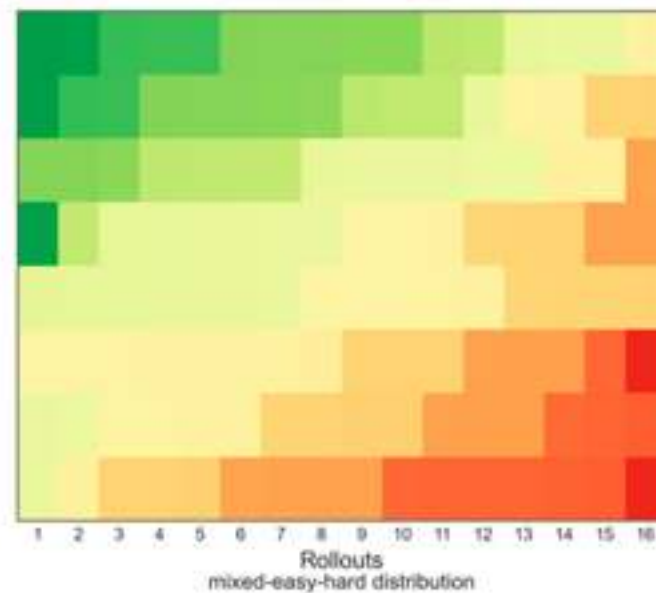
Setup	Performance	MI Proxy	Entropy	Stable
No filter (baseline)	0.17	0.54	2.76	✗
RV filter	0.38 (+0.20)	0.84 (+0.29)	1.64	✓
Entropy filter	0.20 (+0.02)	0.41 (-0.14)	2.20	✗
Length filter	0.16 (-0.02)	0.91 (+0.36)	1.65	✗

Reward Variance is the only filter to improve
BOTH performance AND Mutual Information.

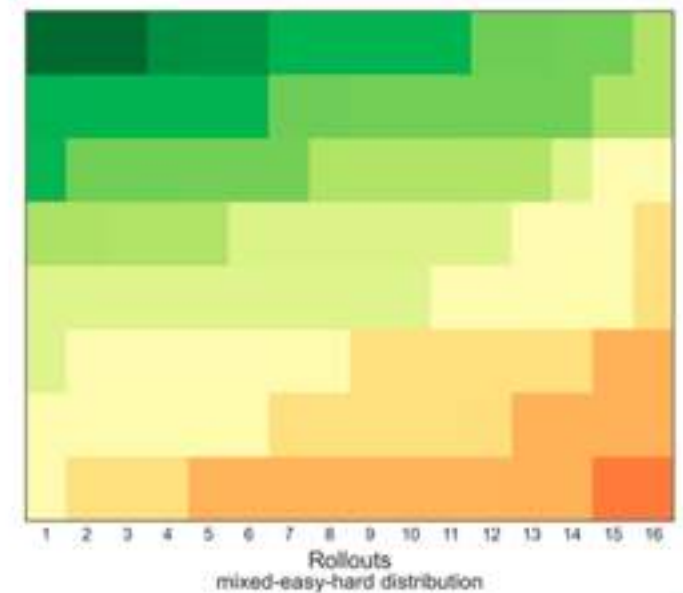
Healthy Reward Distribution along early/mid/late training



Mixed: 2 (25%) Hard: 6 (75%)



Mixed: 5 (62%) Hard: 3 (38%)



Mixed: 5 (62%) Easy: 2 (25%) Hard: 1 (12%)

Learning reasoning through RL

1. What are *failure modes* of RL?

2. What data priors should we encourage for exploration?

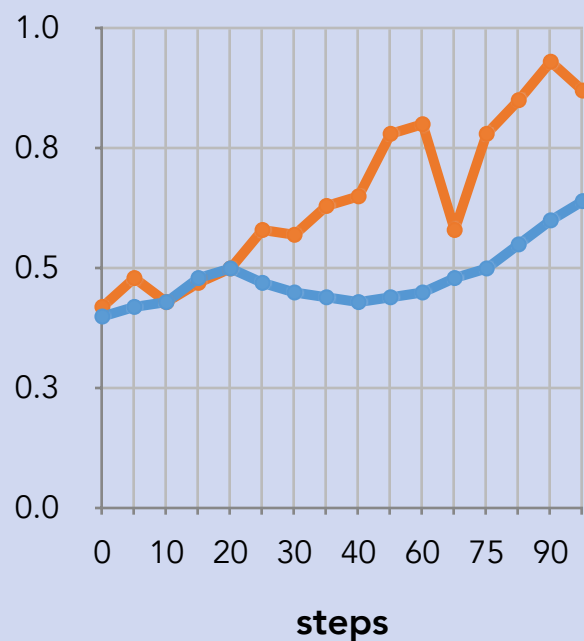
3. When is the best stage to encourage it?

Exploitation vs Exploration

"familiar" env



Webshop

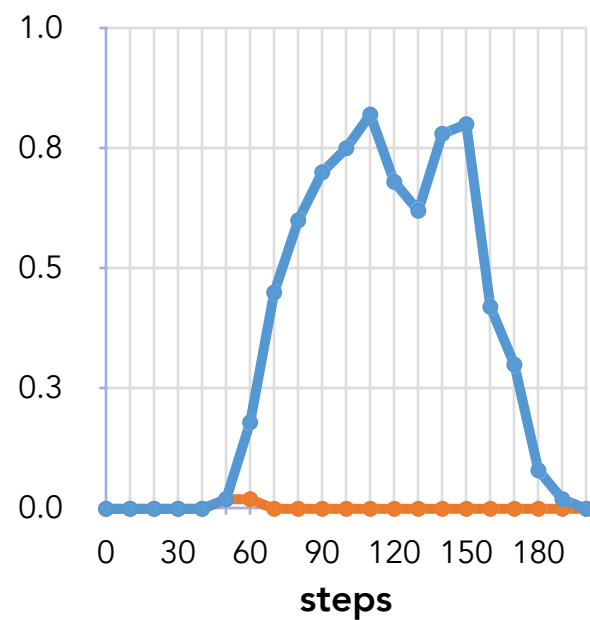
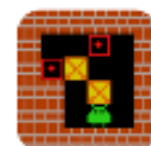


"unfamiliar" env

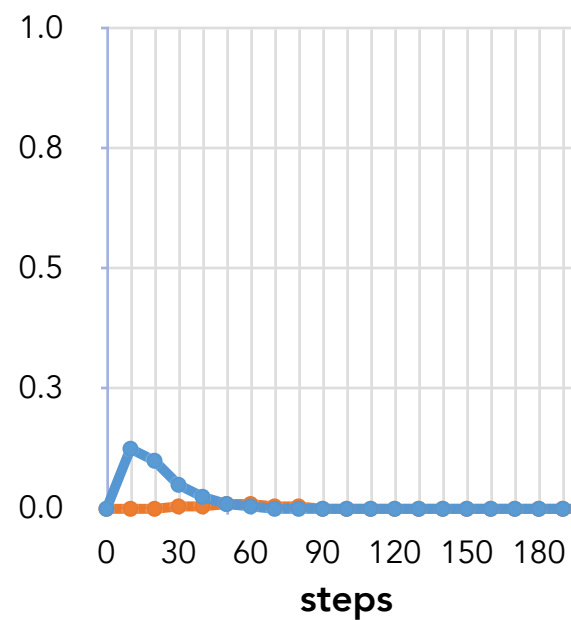
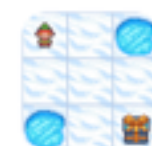
Apply Reinforcement Learning to LLMs ~~×~~ *collapse*



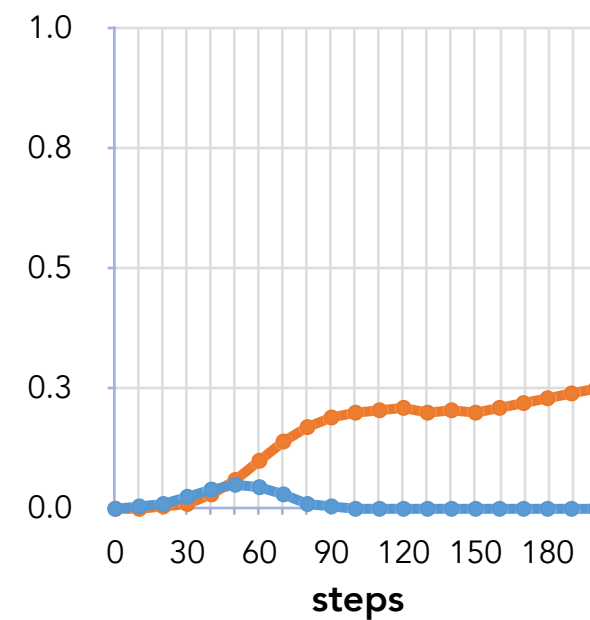
Bandit



Sokoban



FrozenLake



— PPO — GRPO

How to identify OOD environments?

Perplexity of the state space, can be used to indicate **OOD environments**.

Task	Perplexity	
Sokoban	163.9	}
Frozen Lake	187.1	
ALFWorld	6	}
WebShop	11.7	

Unfamiliar

Familiar

Exploitation vs Exploration

Repeat one solution

Exploitation vs Exploration

Alternative solutions

*Especially in **OOD** environments*

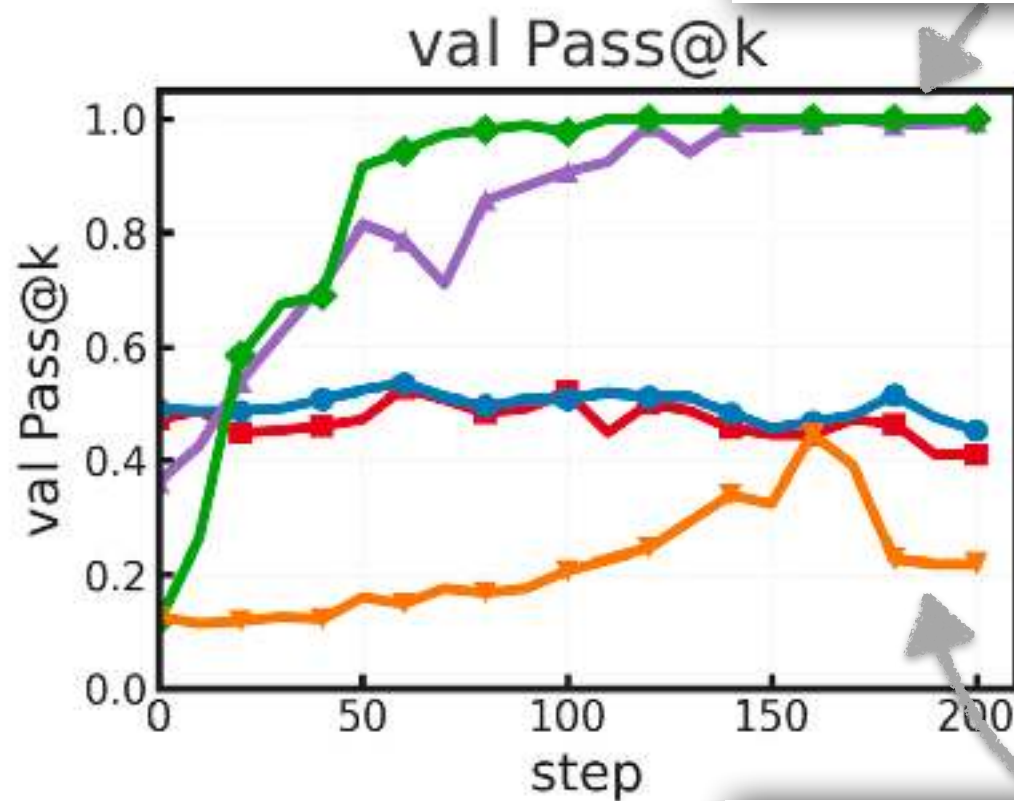
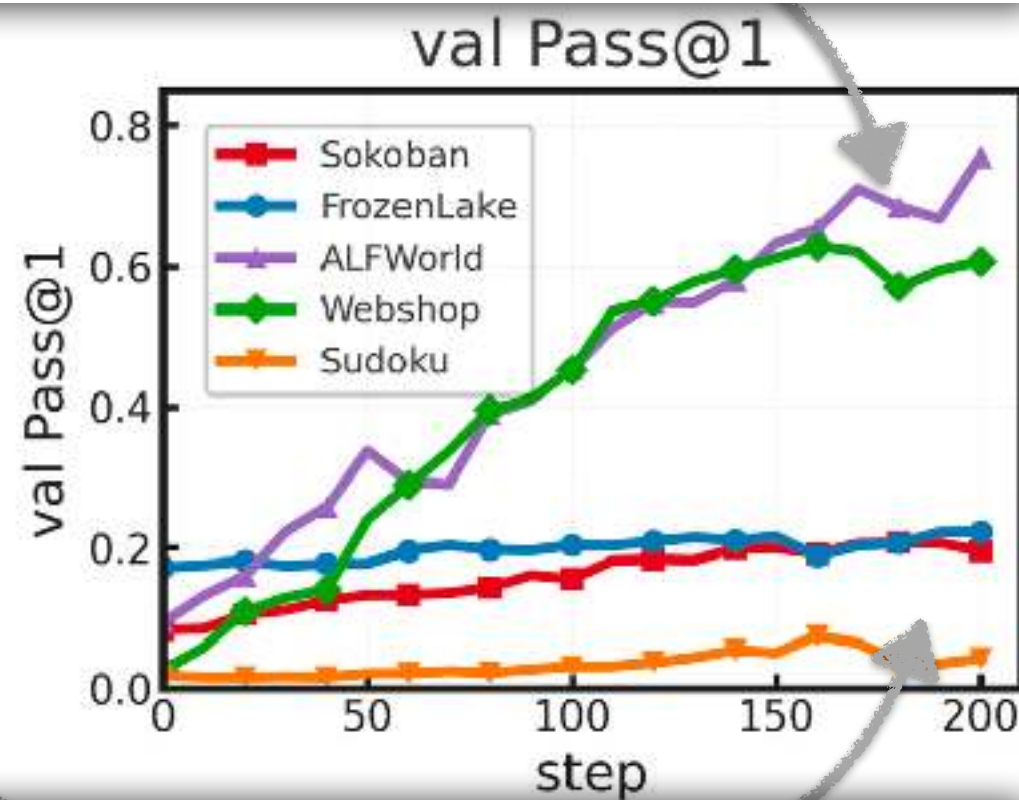
Exploitation vs Exploration

Pass@1

Pass@K

Exploitation vs Exploration

Familiar Environments: Pass@K improves



OOD Environments: Pass@K struggles

Learning **agentic reasoning** through RL

1. What are **failure modes** of RL?

2. What data priors should we encourage for exploration?

3. When is the best stage to encourage it?

Learning **agentic reasoning** through RL

1. What are failure modes of RL?

→ **Exploration** vs *Exploitation*

2. What data priors should we encourage for exploration?

3. When is the best stage to encourage it?

Learning **agentic reasoning** through RL

1. *What are failure modes of RL?*

→ **Exploration** vs *Exploitation*

2. *What **data priors** should we encourage for exploration?*

3. *When is the best stage to encourage it?*

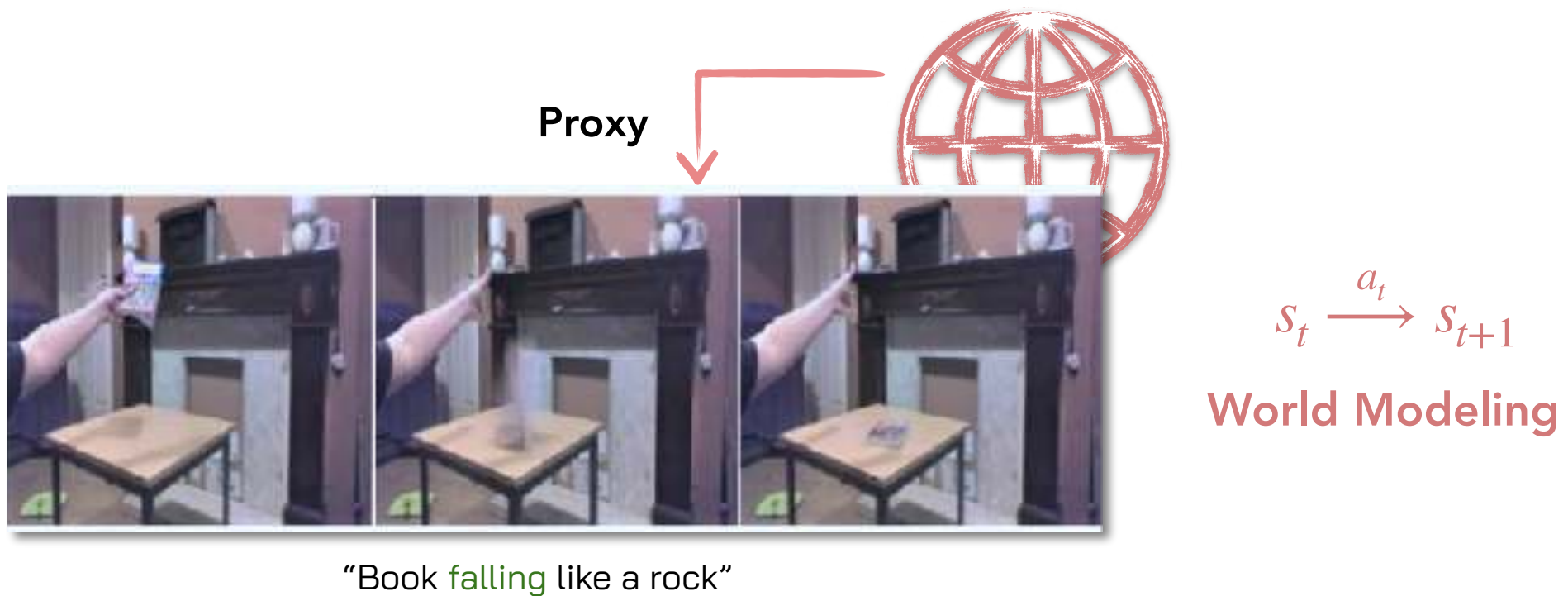
Encourage Exploration:

World Modeling



What is "World Model"?

External World Model



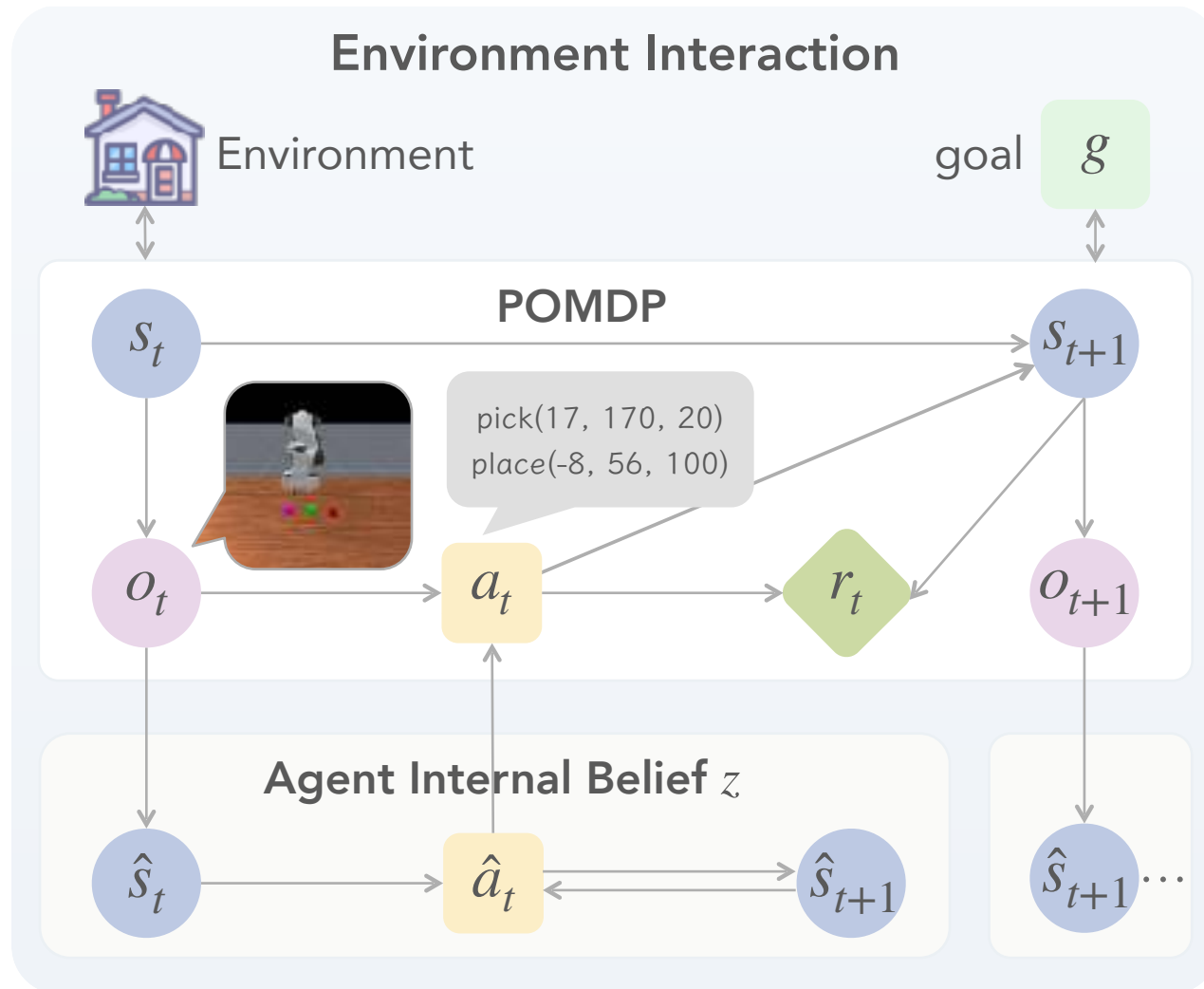
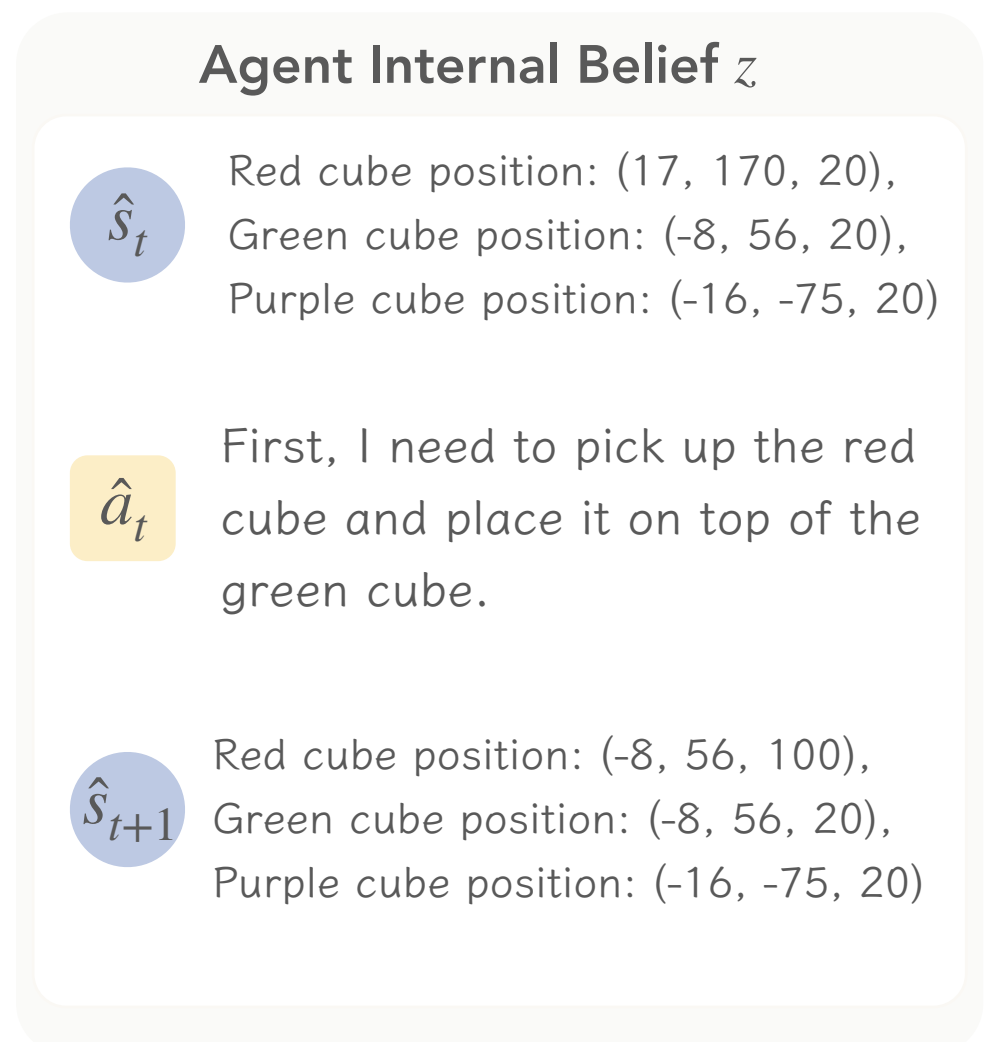
Video: A “Visual Recording” of **World State Changes**

External World Model

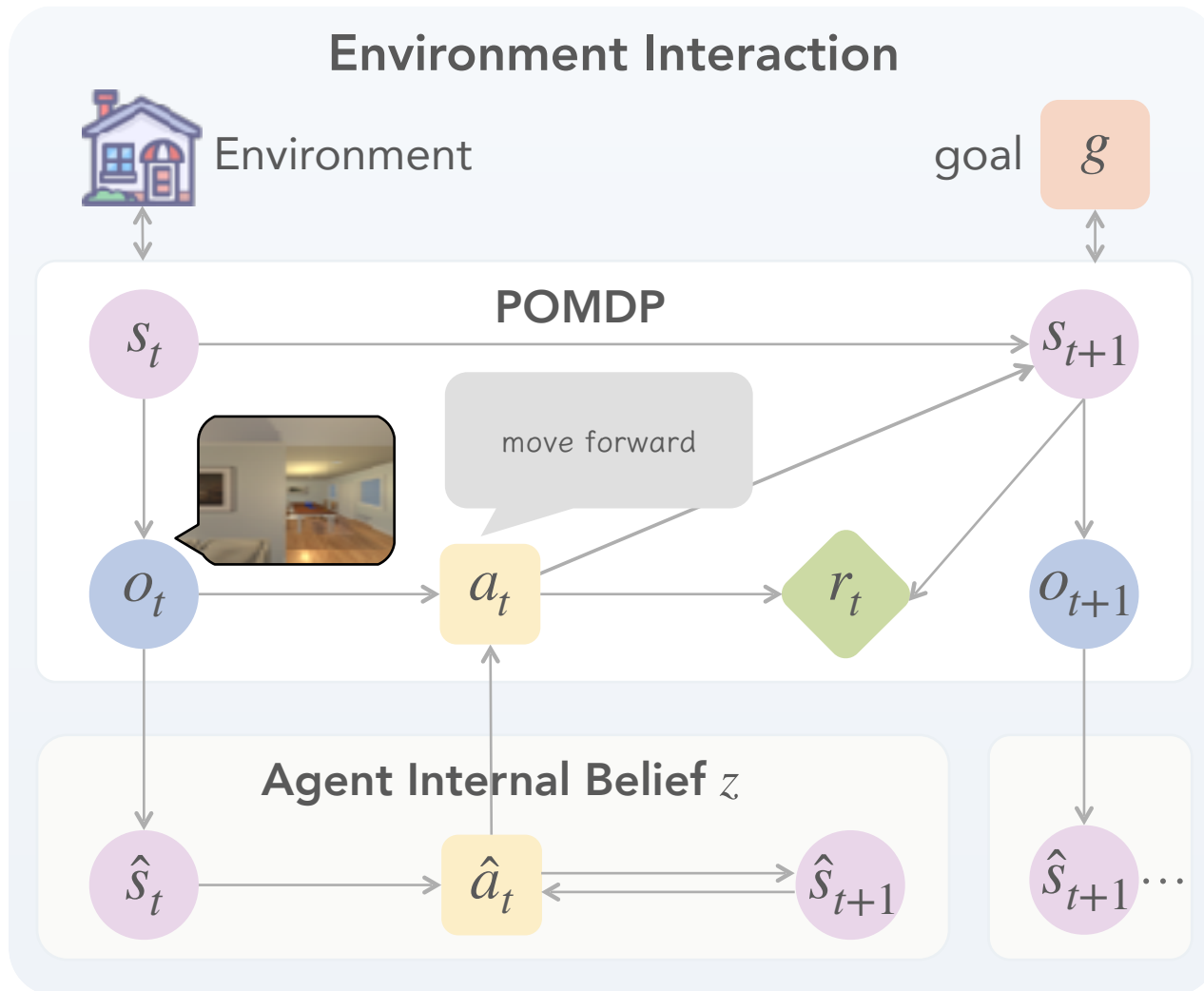
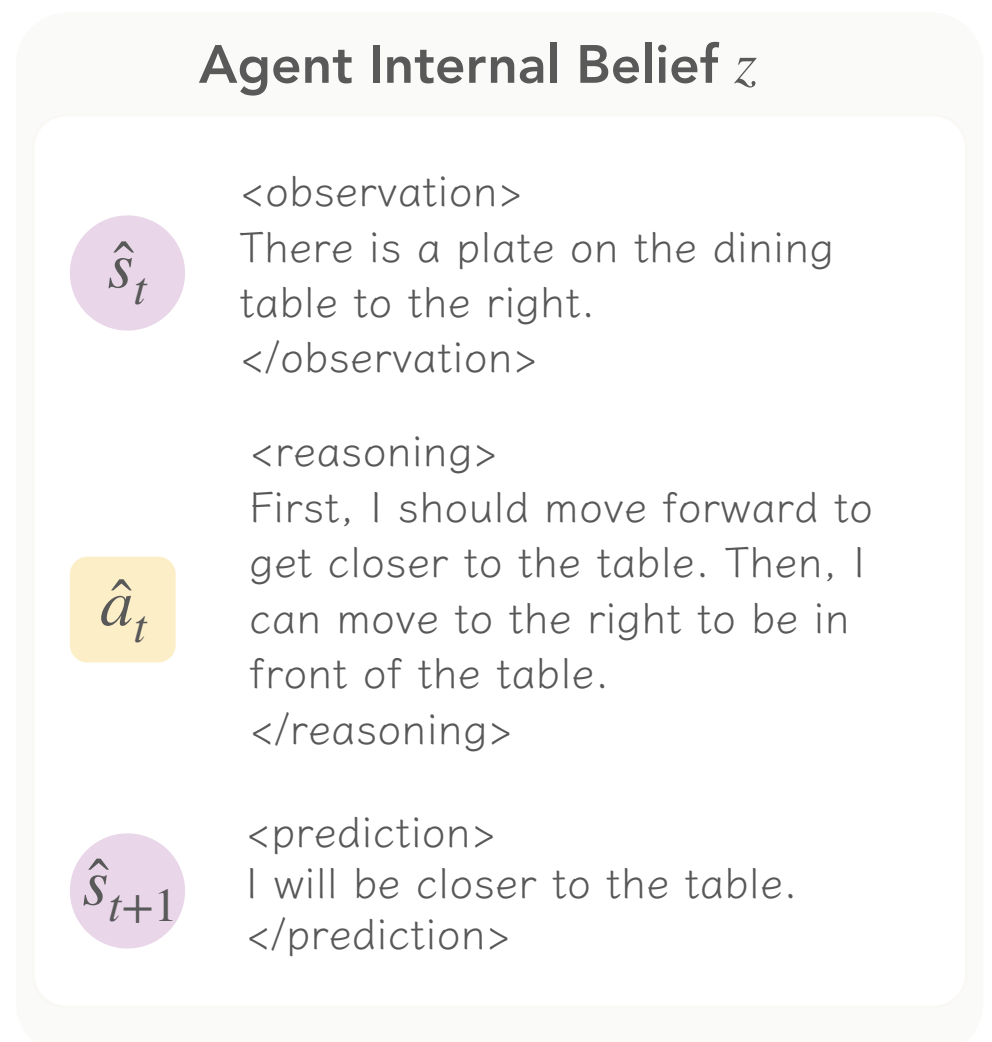


Internal World Model

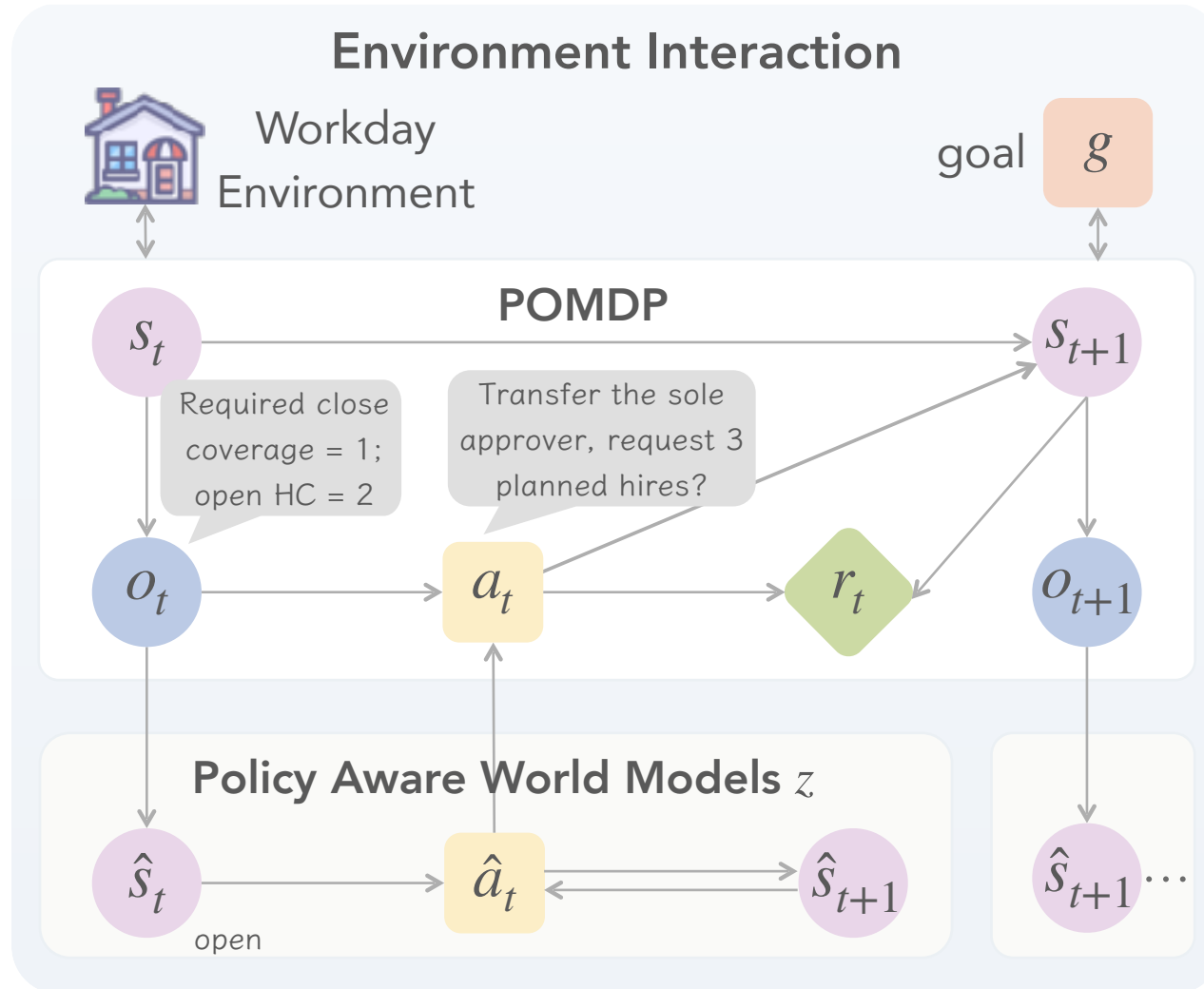
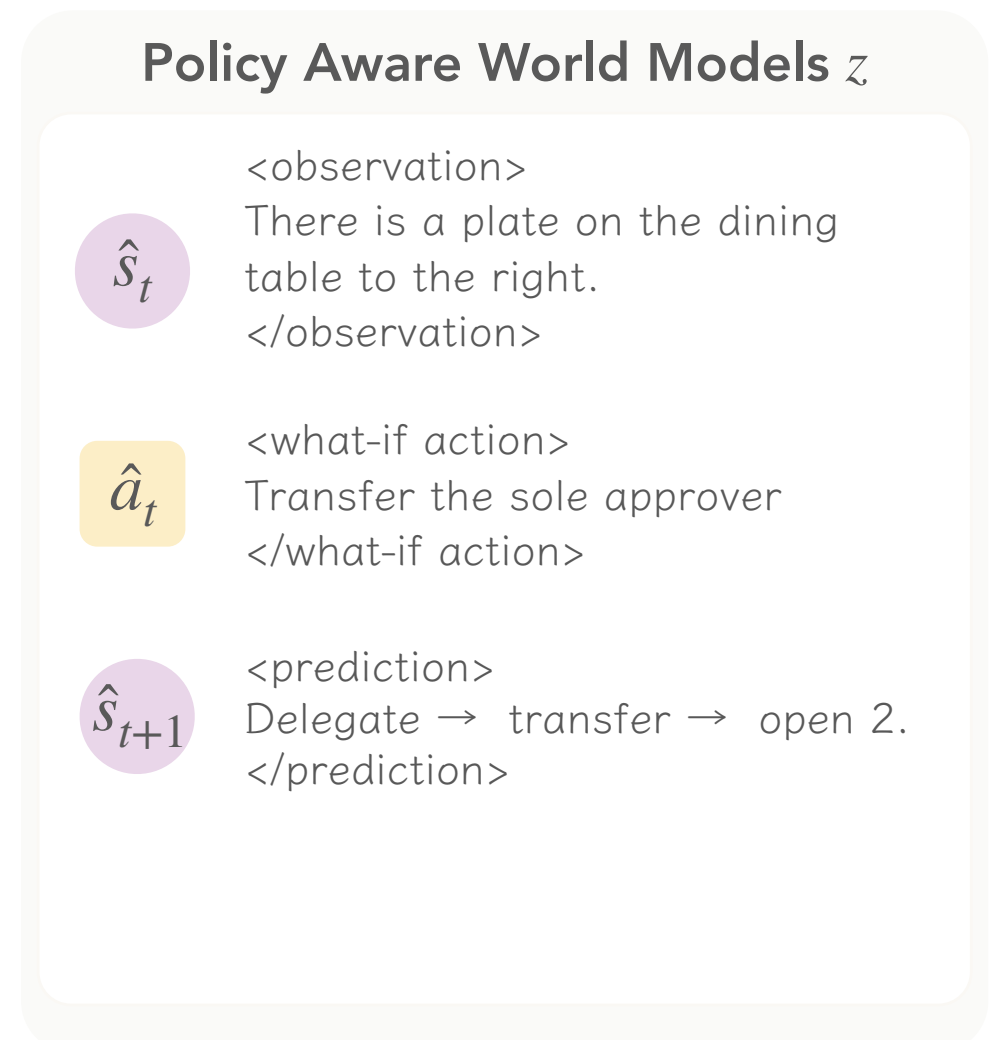
Internal World Model



Internal World Model

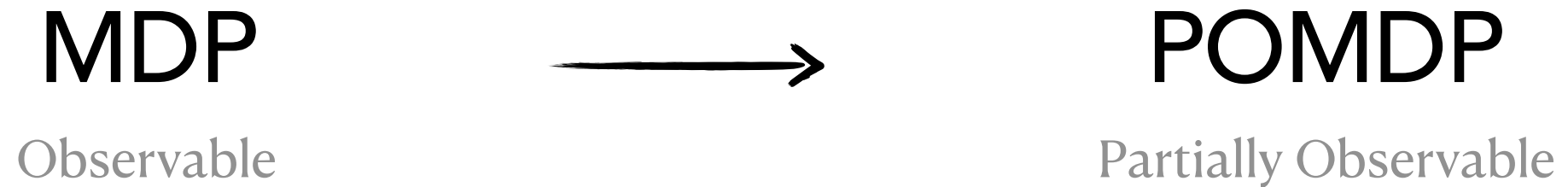


Internal World Model

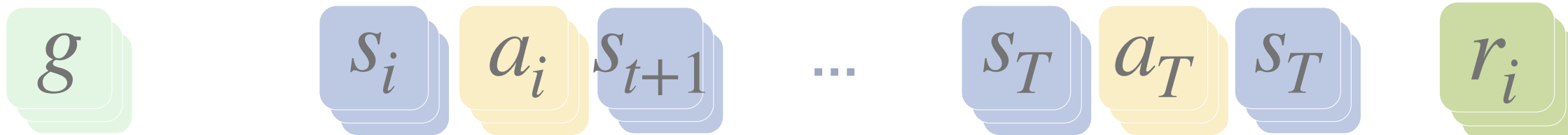




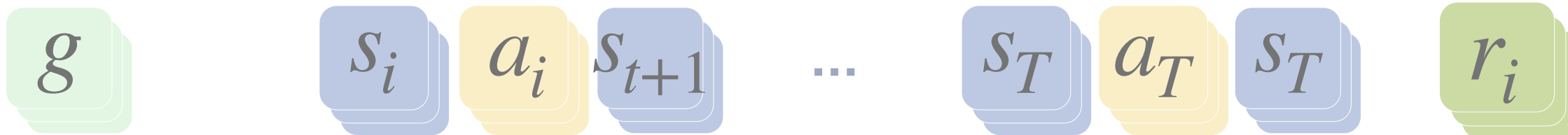
What is "World Model"?



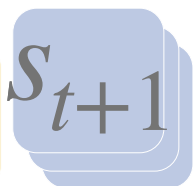
MDP



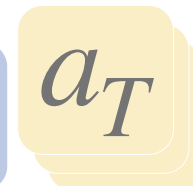
MDP



MDP



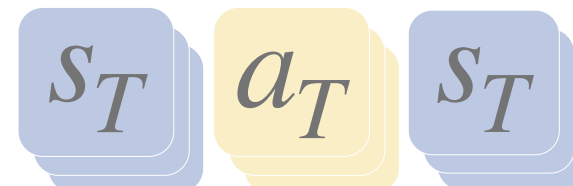
...



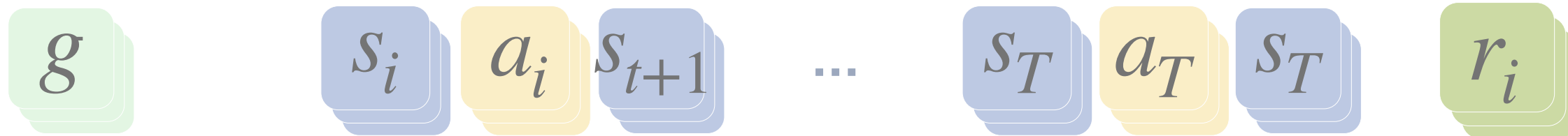
MDP



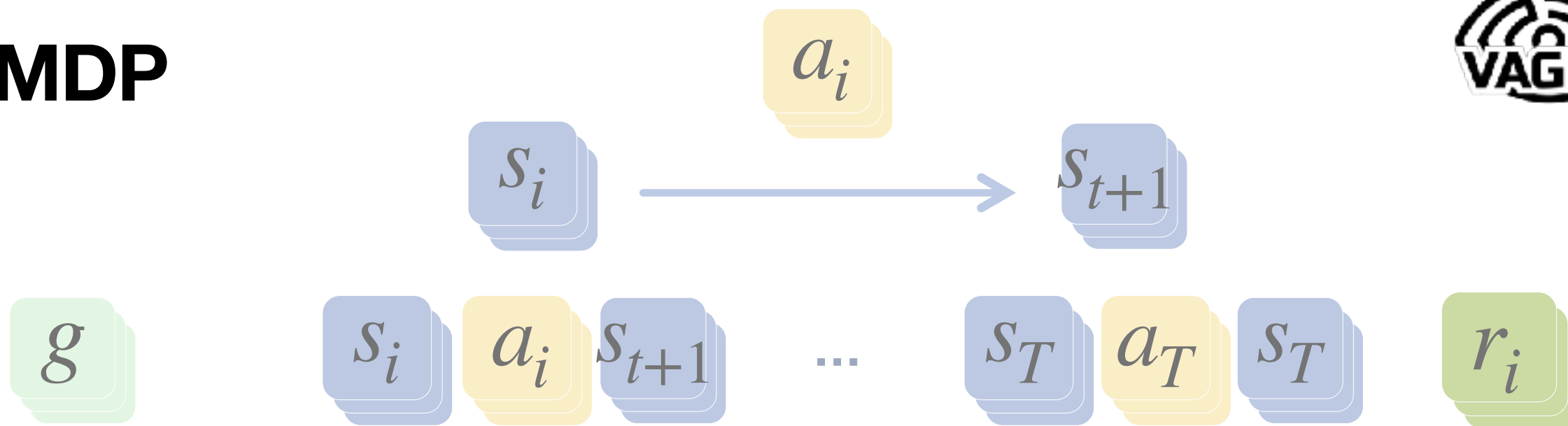
...



MDP



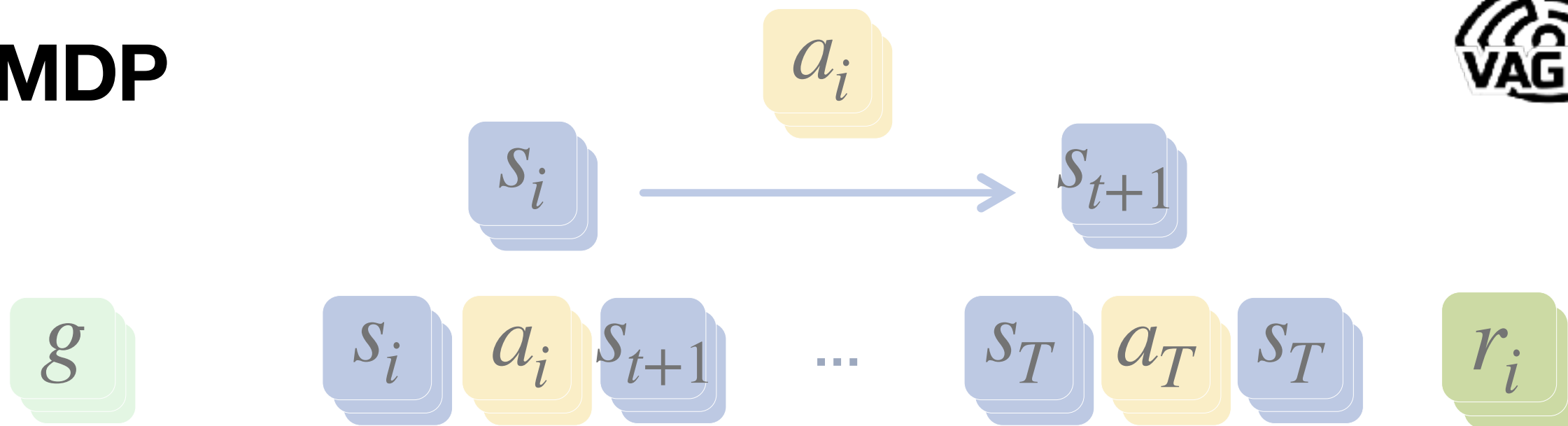
MDP



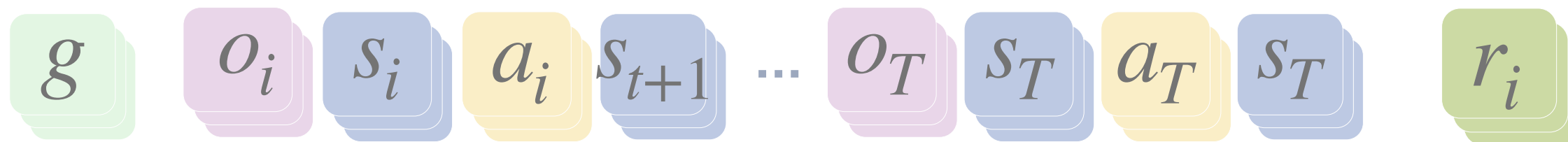
POMDP



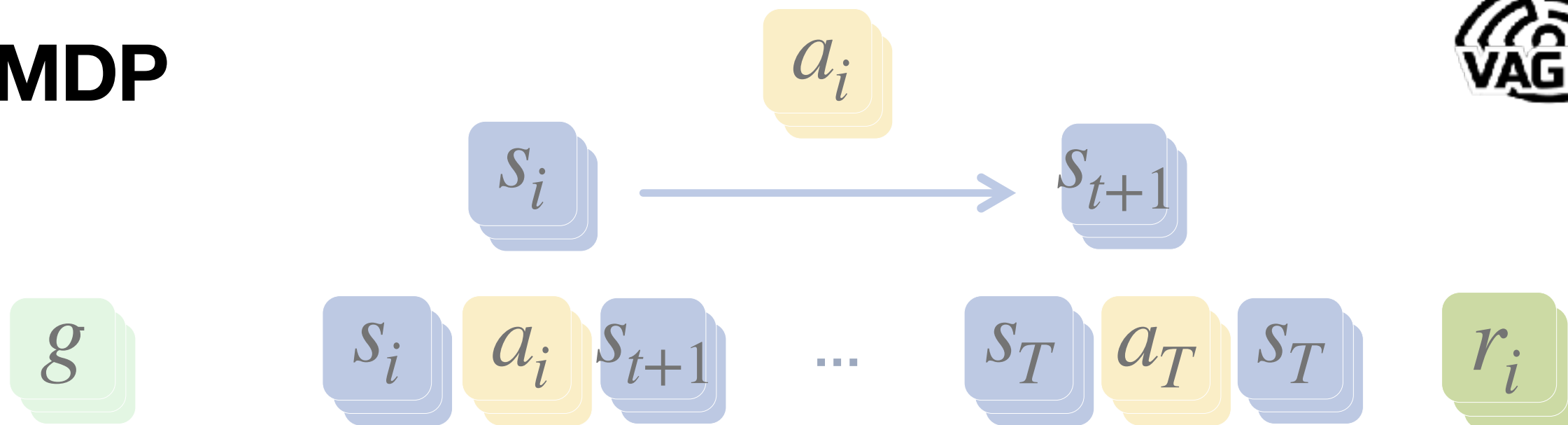
MDP



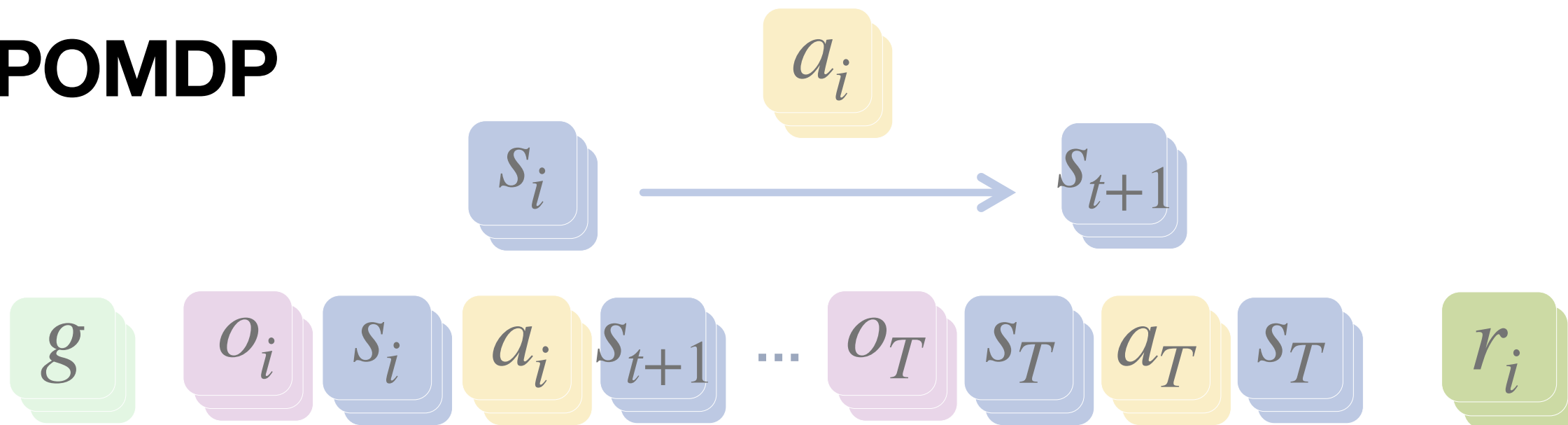
POMDP



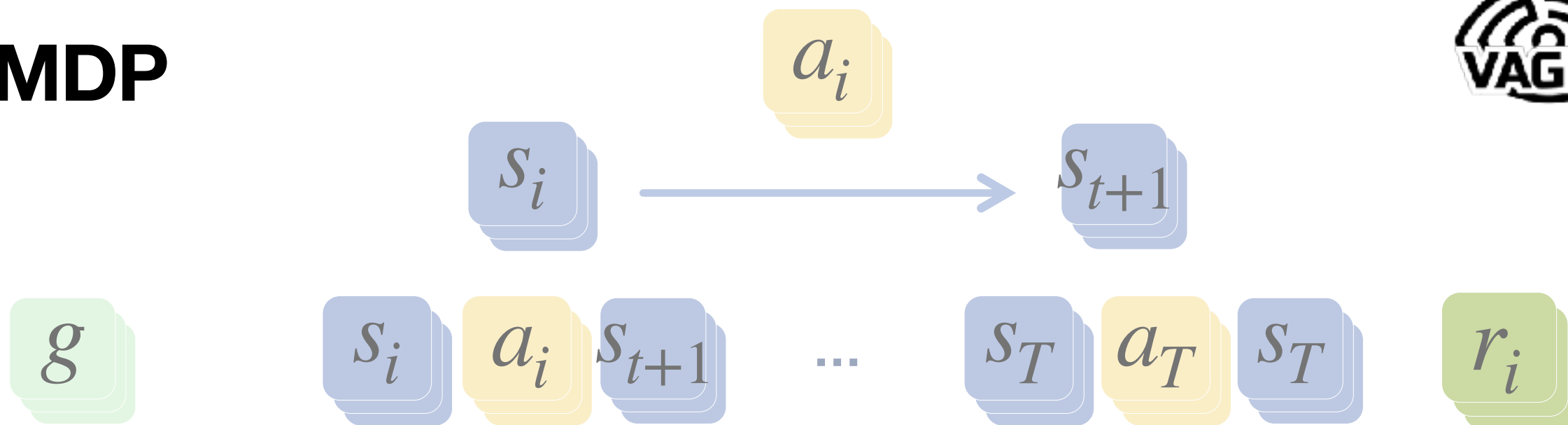
MDP



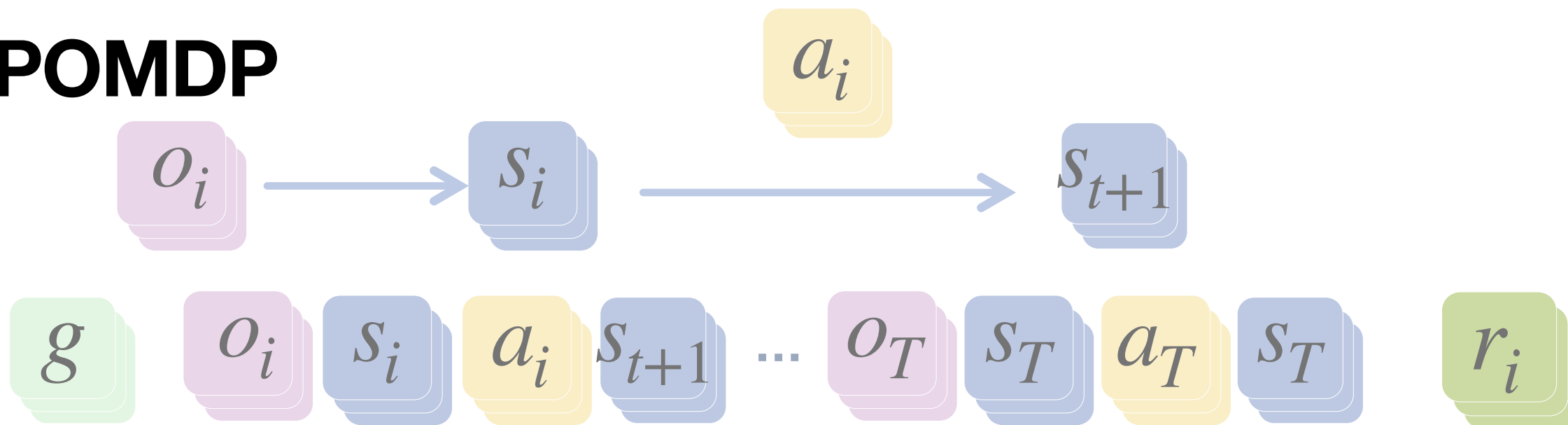
POMDP



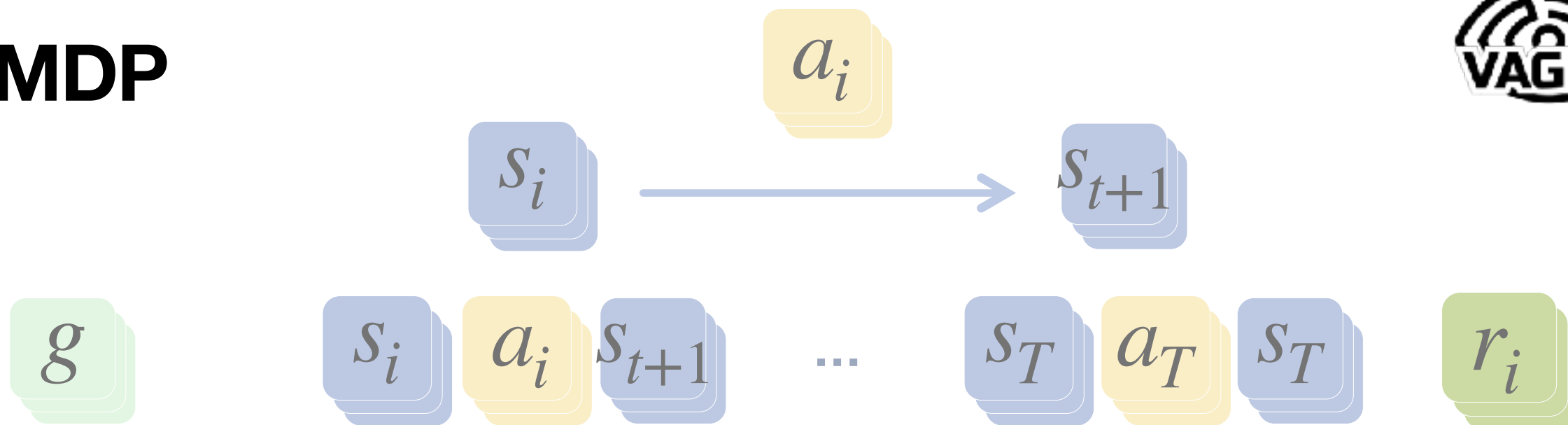
MDP



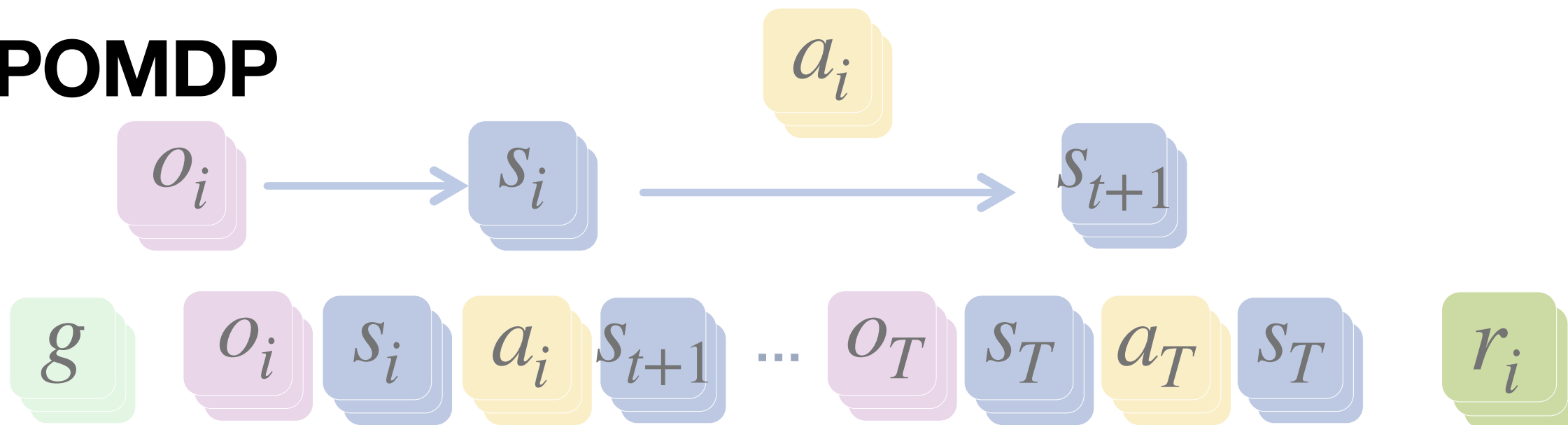
POMDP



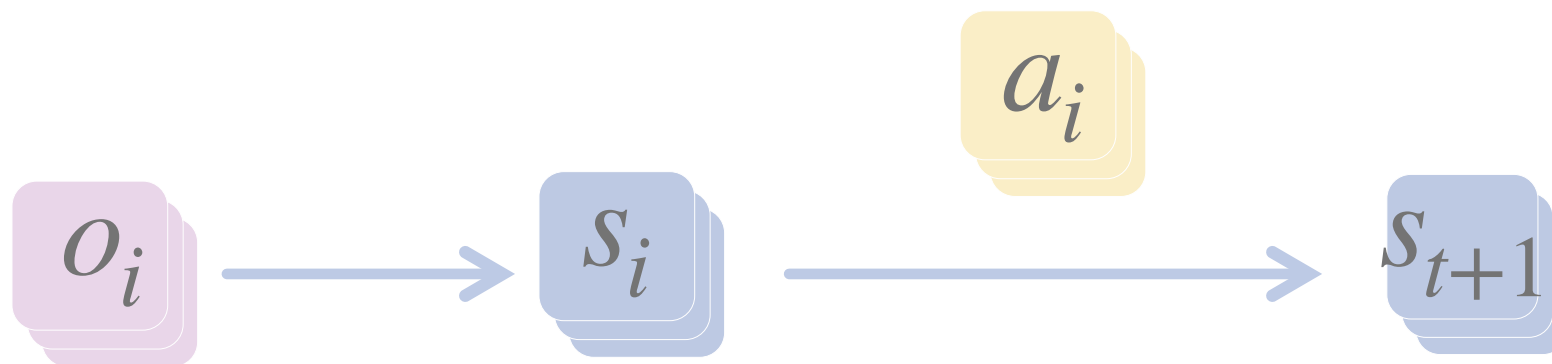
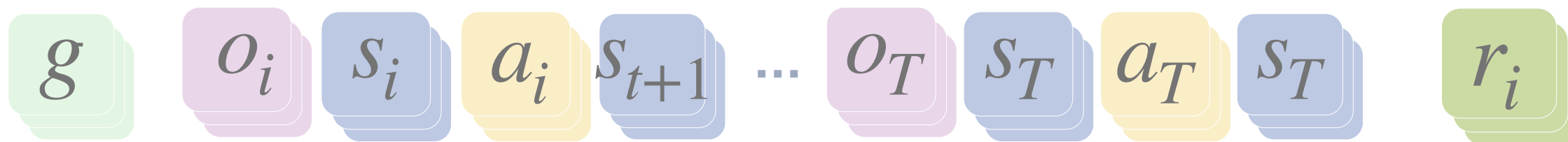
MDP



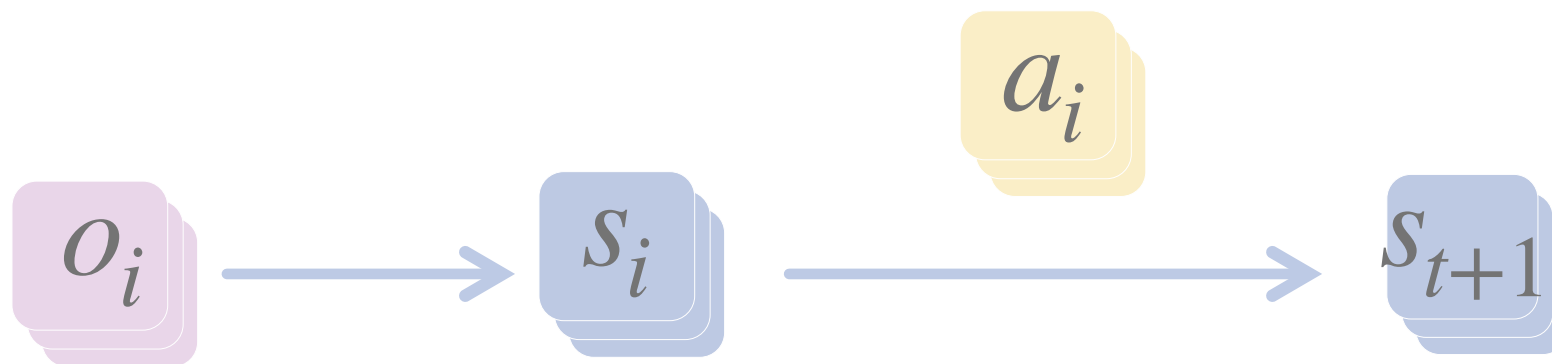
POMDP



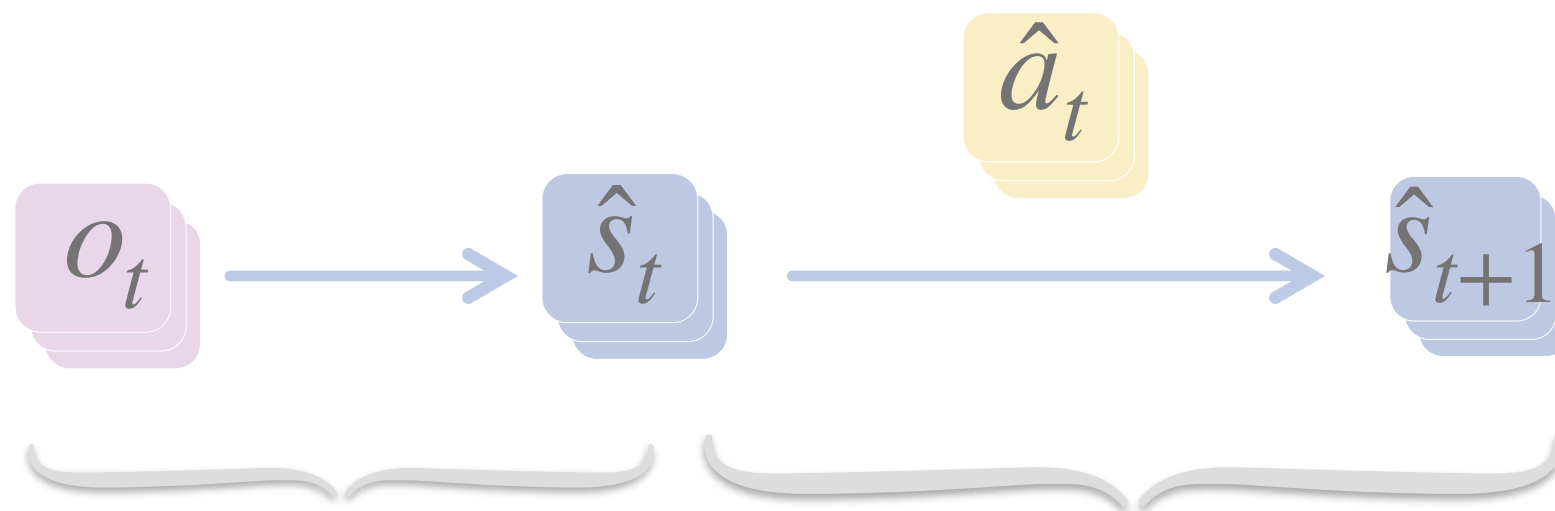
MDP $\rightarrow$ POMDP



World Modeling



World Modeling



State Estimation

$$P(s_t | o_t)$$

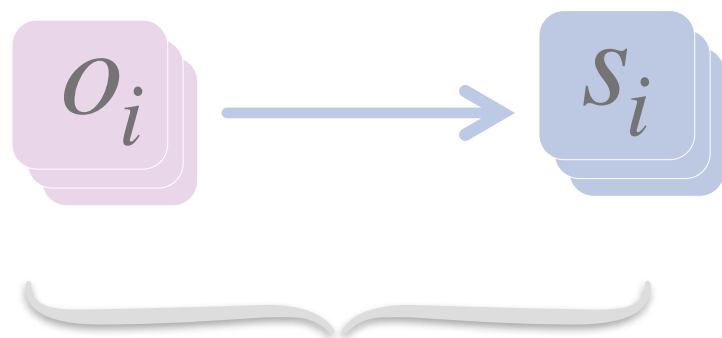
Transition Modeling

$$P(s_{t+1} | s_t)$$



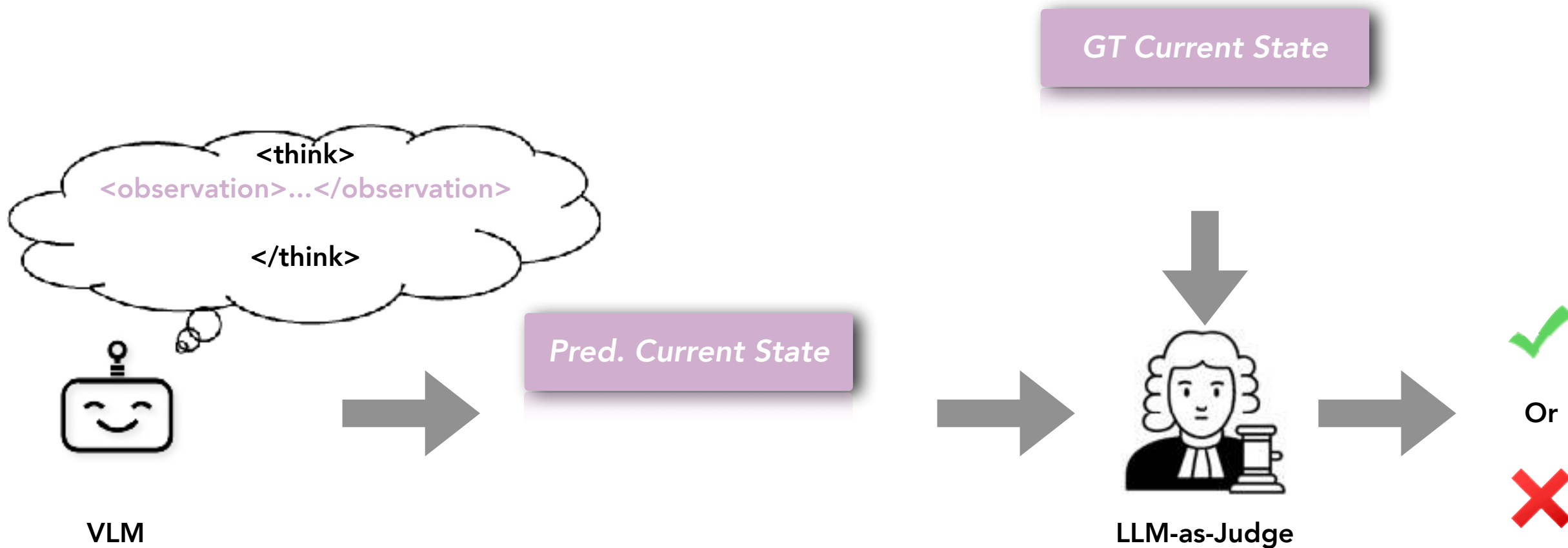
Can **World Modeling** help **Reward Shaping** in RL?

Internal World Model

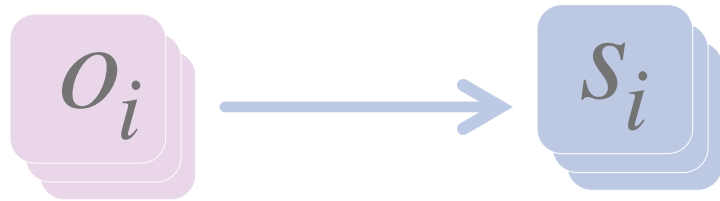


State Estimation

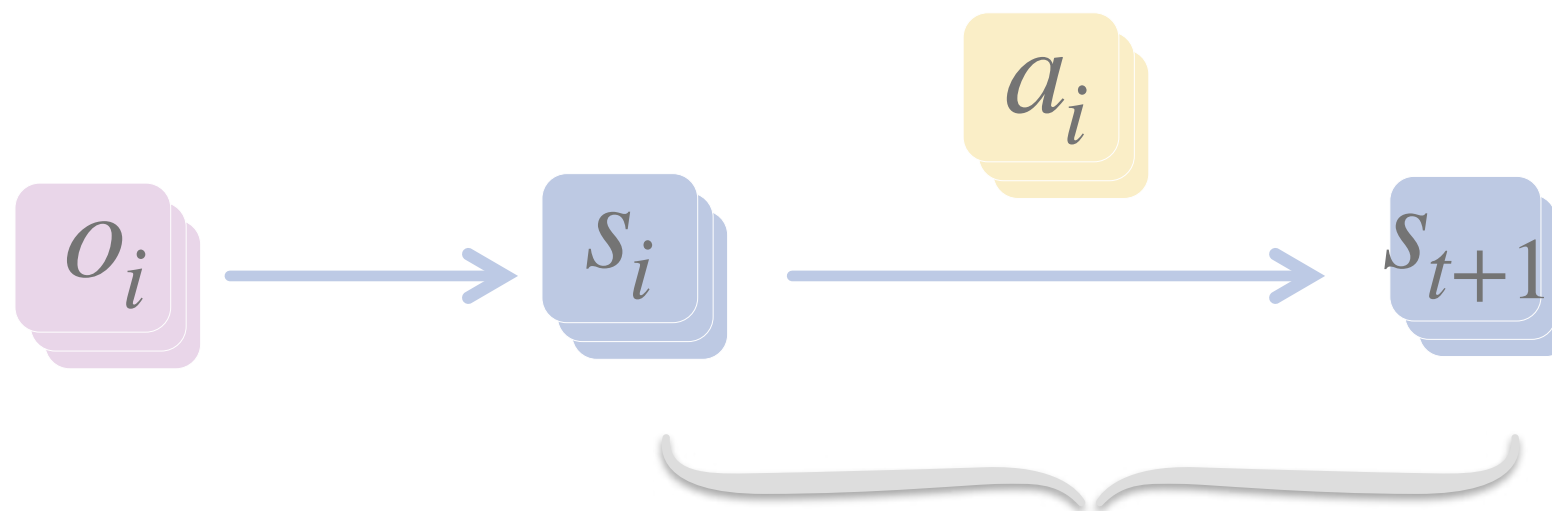
$$P(s_t | o_t)$$



Internal World Model



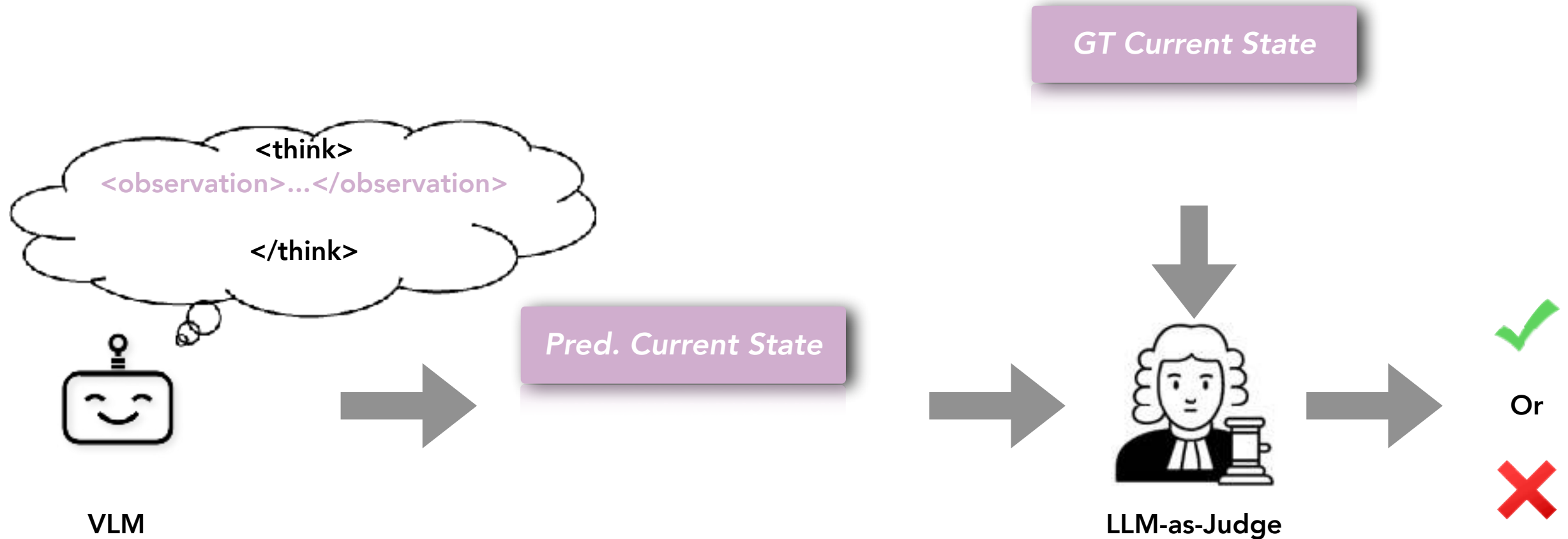
Internal World Model



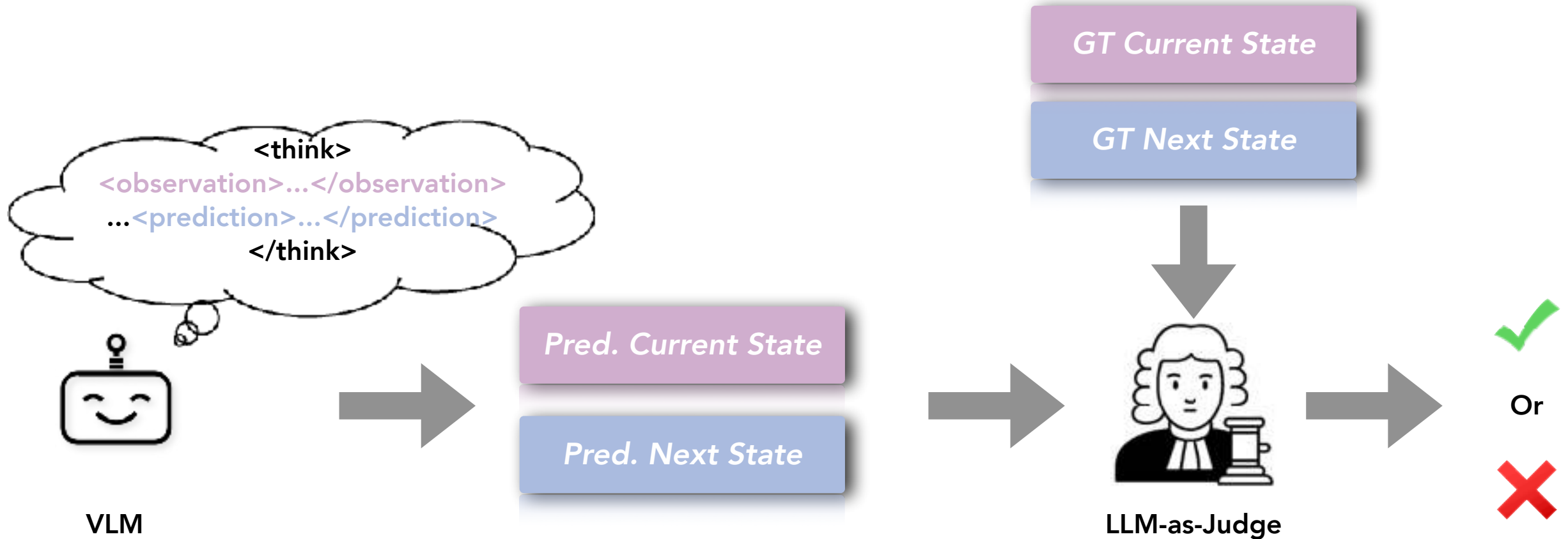
Transition Modeling

$$P(s_{t+1} | s_t)$$

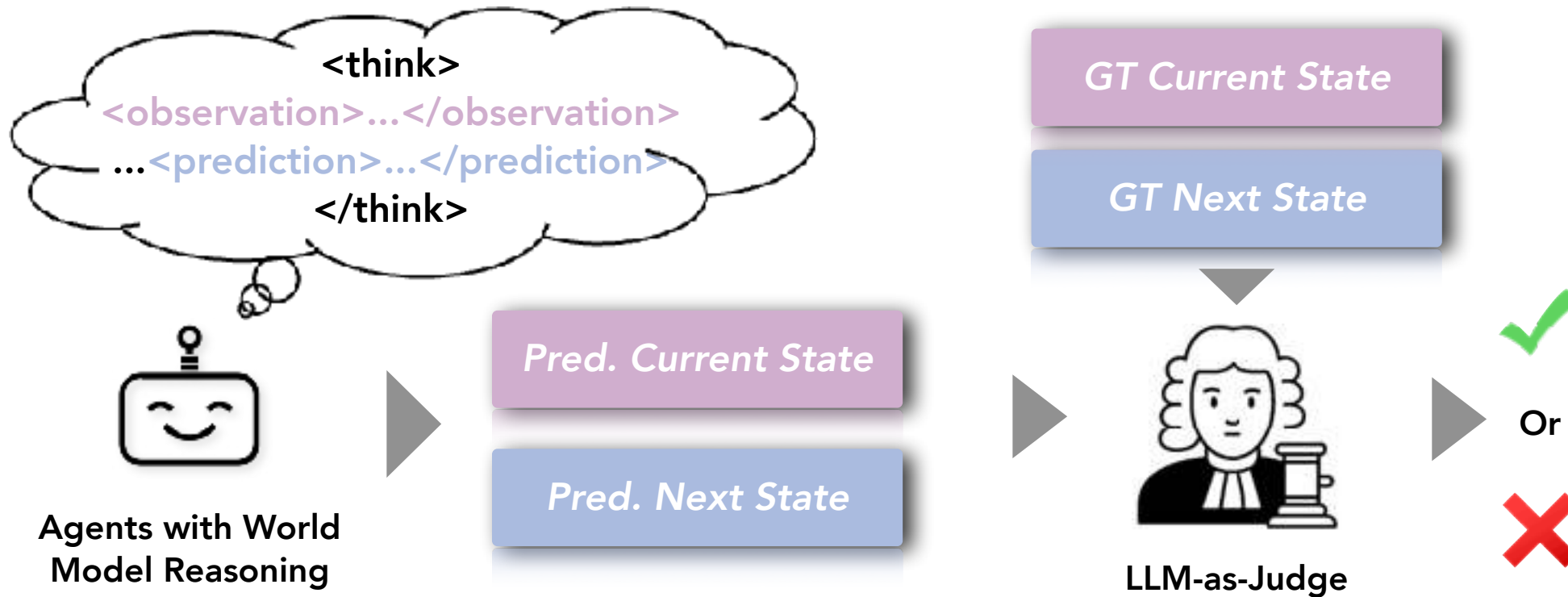
World Model Reasoning Reward



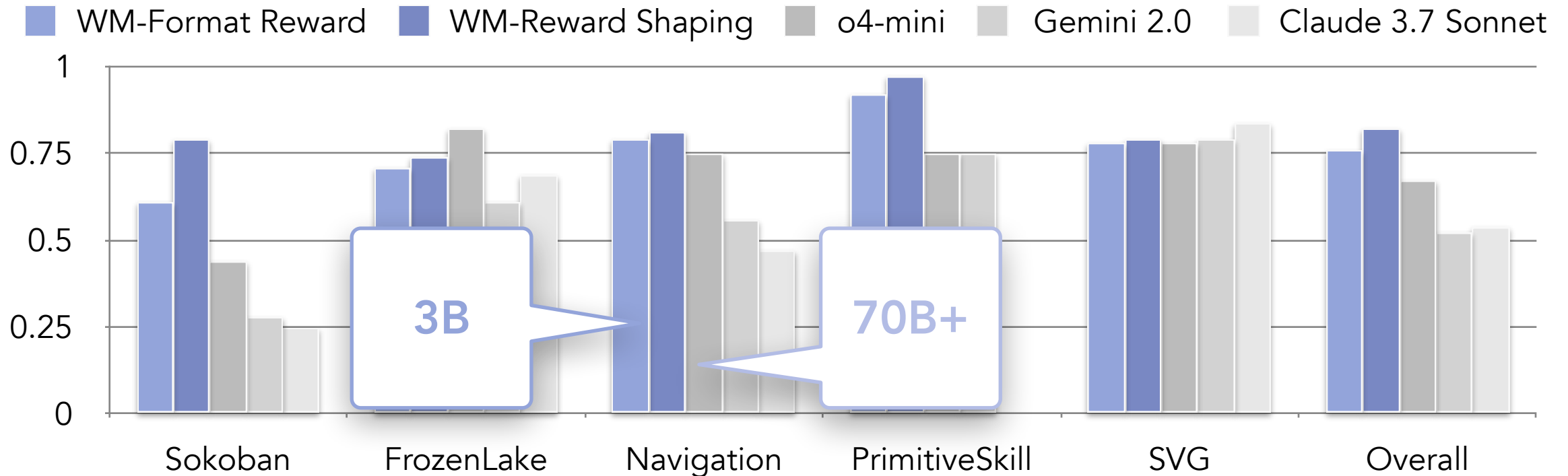
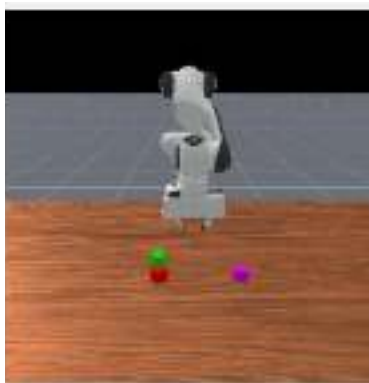
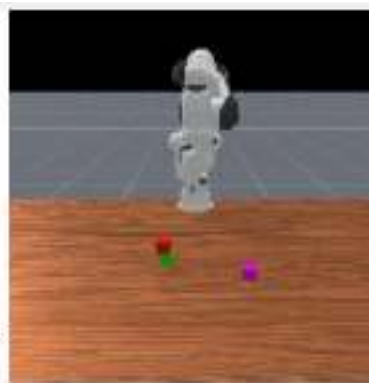
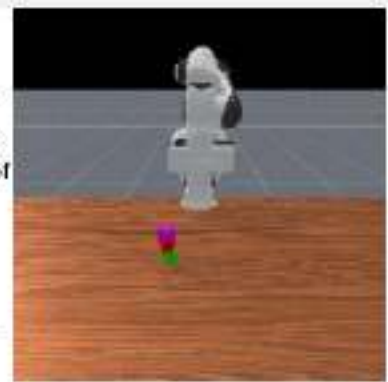
World Model Reasoning Reward



World Model Reasoning Reward



World Model helps



Learning reasoning through RL

1. *What are failure modes of RL?*

→ *Exploration vs Exploitation*

2. *What priors should we encourage for exploration?*

→ *World Model*

3. *When is the best stage to encourage it?*

When a child enters an unfamiliar room...

Goal: find a toy



Task Data

Play Data

STAGES

Pre-training

Self-Supervised
Finetuning

Mid-training

Reinforcement
Learning

DATA

g

a_i

a_T

annotated solution

s_i

a_i

s_{t+1}

...

s_T

a_T

s_T

no annotation

g

s_i

a_i

s_{t+1}

...

s_T

a_T

s_T

r_i

reasoning about solution

Learning reasoning through RL

1. *What are failure modes of RL?*

→ *Exploration vs Exploitation*

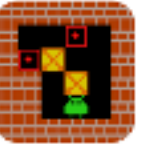
2. *What priors should we encourage for exploration?*

→ *World Model*

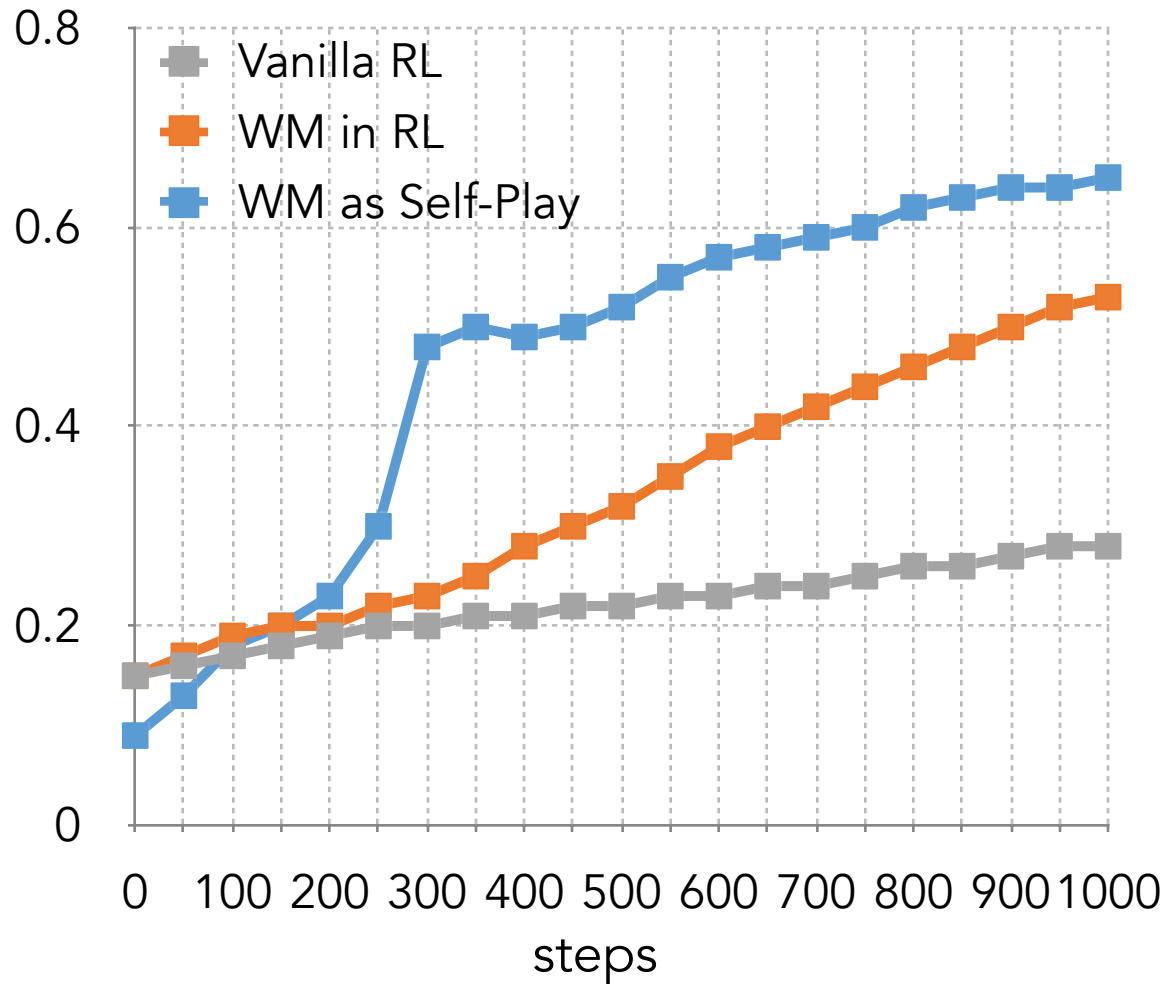
3. *When is the best stage to encourage it?*

→ *Mid-training vs RL*

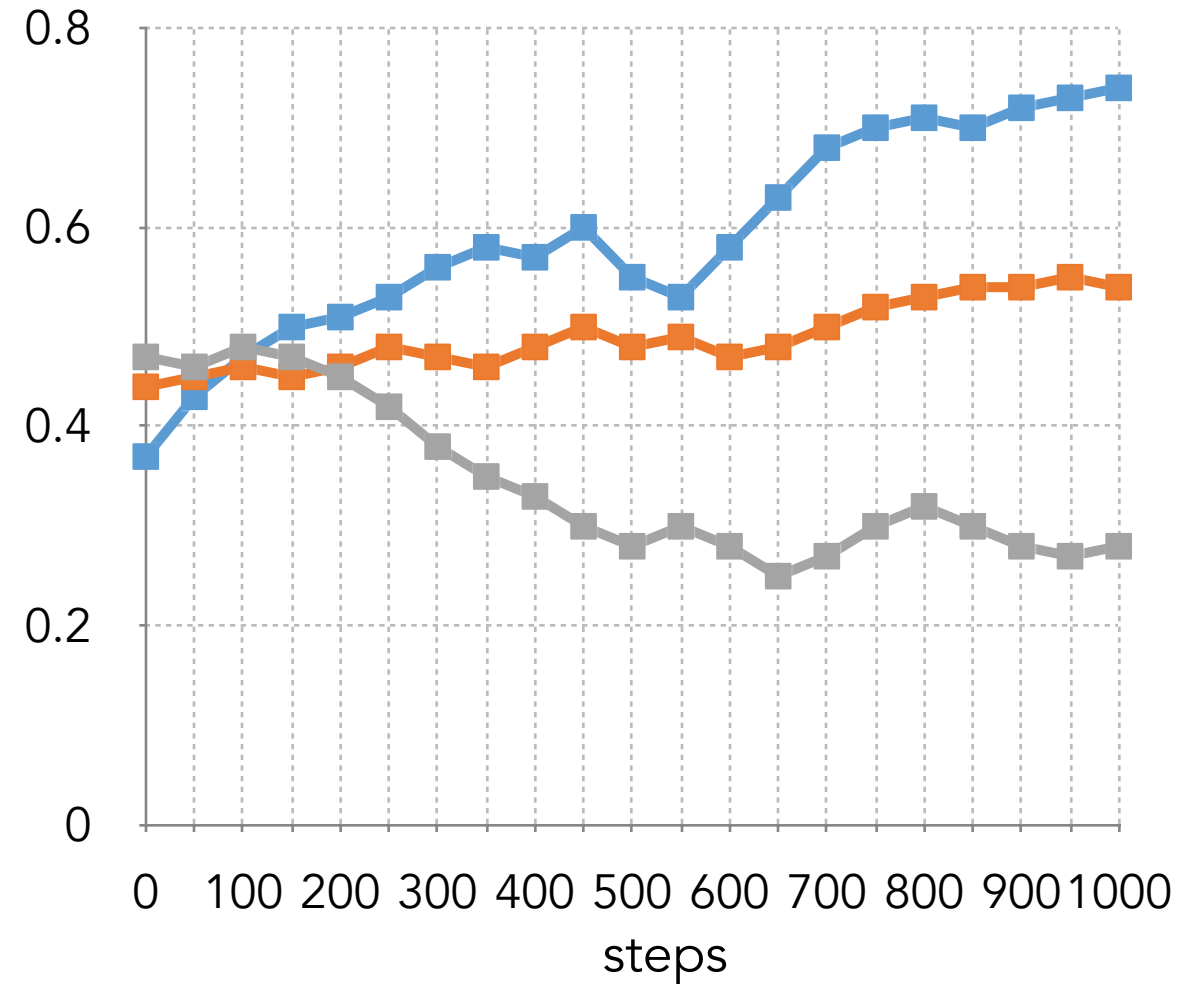
Exploration vs Exploitation



Pass@1



Pass@K



PRIORS

Common
Knowledge

STAGES

Pre-training

Self-Supervised
Finetuning

Mid-training

Reinforcement
Learning

DATA

Static Internet Data:
Books, Images, Videos...

Human Annotations:
<instructions, answers>

Play Data:
Env, Tools, External Data

Task Feedback:
Verifiers (Goal Definitions)

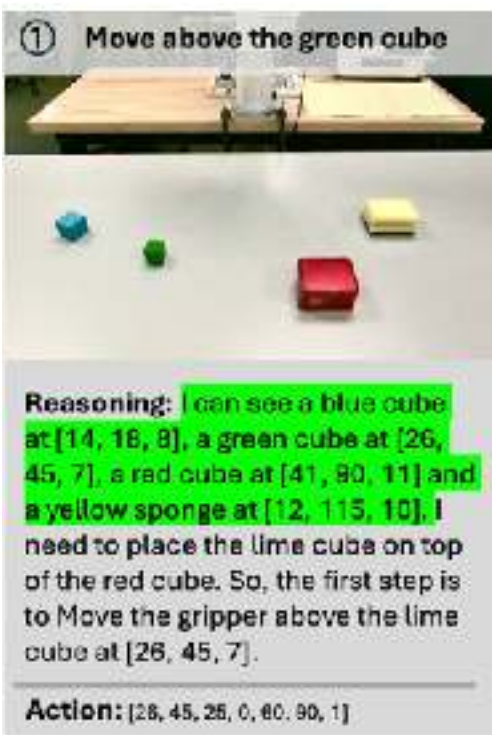
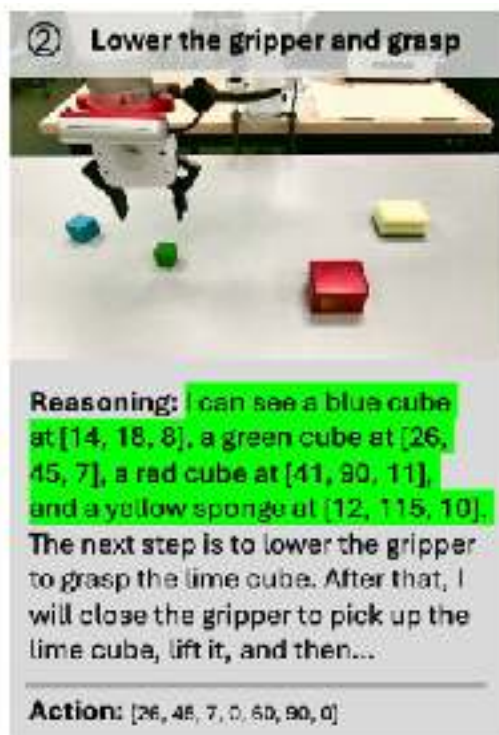
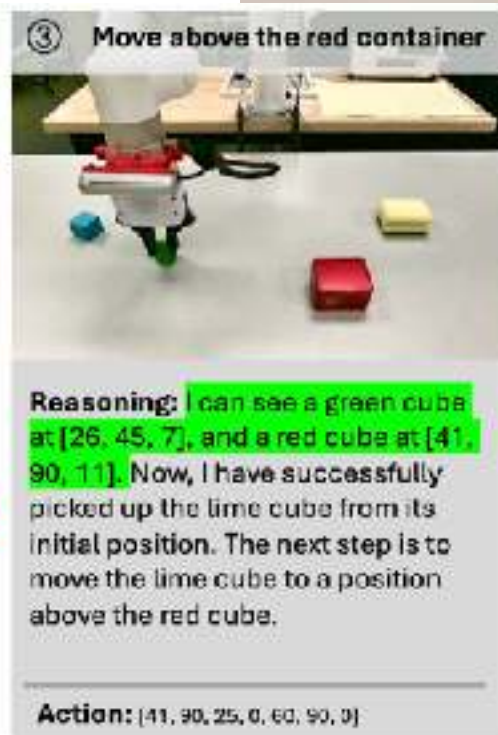
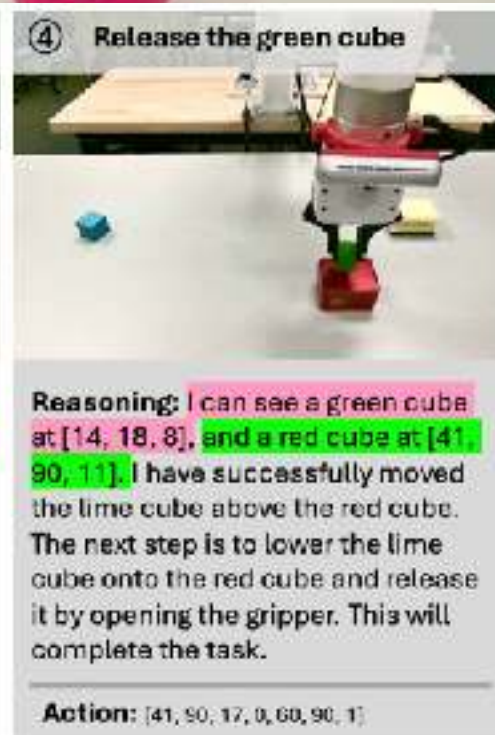
Brute-force Scaling



Smart Scaling

Algorithms for the win!

Embodied Reasoning Agent via VLM Reasoning



RAGEN: Impact

Fork 209 Starred 2.5k

Awesome Work Powered or Inspired by RAGEN

- [ROLL](#): Efficient Scaling Library for RL with LLMs 2k Stars
- [VAGEN](#): Training Visual Agents with multi-turn RL 431 Stars
- [Search-R1](#): Train LLMs to reason and call a search engine with RL 4.2k Stars
- [ZeroSearch](#): Incentivize LLM search capability without searching 1.3k Stars
- [Agent-R1](#): Training Powerful LLM Agents with End-to-End RL
- [OpenManus-R1](#): RL tuning for LLM agents 39k Stars
- [MetaSpatial](#): Reinforcing 3D Spatial Reasoning in VLMs 202 Stars
- [e3](#): Efficient Yet Effective Search Agent Training via RL

REASONING GYM: Reasoning Environments for Reinforcement Learning with Verifiable Rewards

Zafir Stojanovski^{1*} Oliver Stanley^{1,2*} Joe Sharratt^{1*} Richard Jones^{1,2*}

Abdulkareem Adeboye² Jean Kaddour^{2†} Andreas Köpf²

SPIRAL: SELF-PLAY ON ZERO-SUM GAMES INCENTIVIZES REASONING VIA MULTI-AGENT MULTI-TURN REINFORCEMENT LEARNING

Bo Liu¹, Simon Yue², Zichen Liu^{1,2}, Leon Guenther¹, Pengfei Qi^{1,2}, Hamid Hazzeth², Mickael Luth¹, Chester Tan¹, Weyan Shi², Min Lin², Wei Sun Lee¹, Narasimha Joshi^{2*}

THE DEVIL BEHIND MOLTBOOK: ANTHROPIC SAFETY IS ALWAYS VANISHING IN SELF-EVOLVING AI SOCIETIES

Chenru Wang¹, Chaochao Li^{1*}, Songyang Liu¹, Zejian Chen¹, Jinyu Hou¹, Ji Qi¹, Bai Li¹, Liliang Zhang¹, Qiwei Ye^{2*}, Zheng Lin², Xu Chen², Xi Zhang¹, Philip S. Yu¹

WebThinker: Empowering Large Reasoning Models with Deep Research Capability

Xiaoli Li¹, Jinjie Jin^{1*}, Guanting Dong^{1*}, Honglin Qian², Yongkang Wu¹, Ji-Rong Wen¹, Yuntao Zhu¹, Zhiheng Dong^{1†}

Search-R1

Rowen Hu
@BorevUnit3

Our pipeline is based on verl (github.com/volcengine/verl), a highly efficient RL framework. We would also like to thank HryZero (github.com/HryZero/TinyLlama) and RAGEN (github.com/ZihanWang314/RAGEN) which inspired this work.

12:52 PM · Feb 26, 2025 · 26,816 Views

2 5 19 26



Tanishq Mathew Abraham, Ph.D. @ScienceLaw · Dec 9, 2024
A new tutorial on RL by Kevin Patrick Murphy, a Research Scientist at Google DeepMind who also wrote several comprehensive, well-regarded textbooks on ML/DL.

This ought to be a good read !!

6 x 5	Algorithm	Example	137
6.2.5.1	Example: RAGEN		137
6.2.5.2	Decline with invalid actions		158
6.2.5.3	Time-step learning		158
6.2.5.4	Self-play for LLM training		159

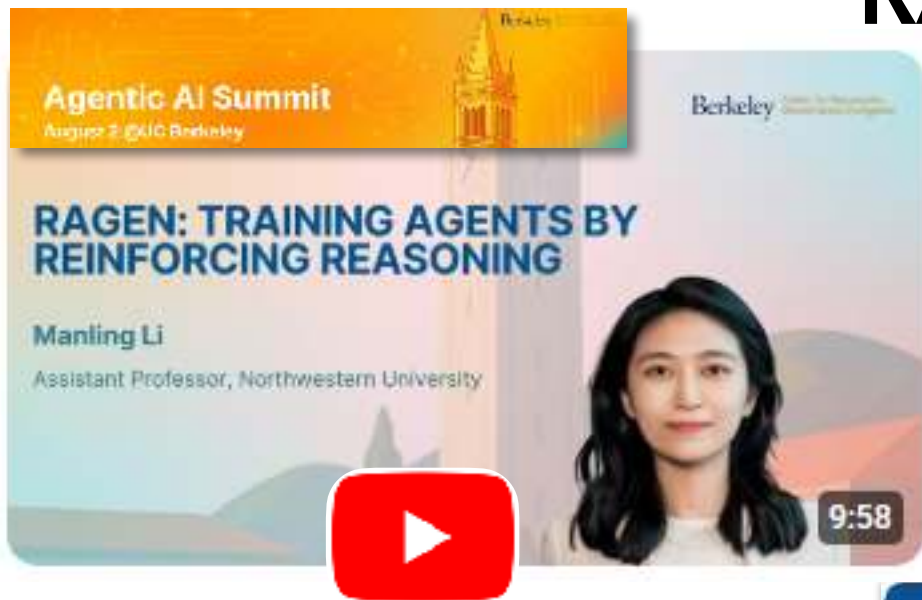
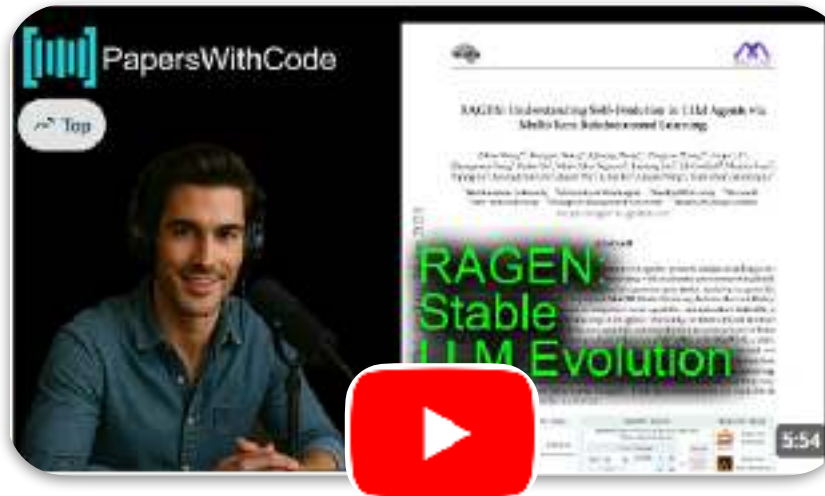


VentureBeat · Apr 23, 2025
Former DeepMind researchers release new method for training reliable AI agents: RAGEN



Wenhu Chen @WenhuChen · Apr 23, 2025
Checked ZihanWang's RAGEN. Lots of work has been done on RAGEN's code base! Nice work!

RAGEN: Impact

CONVICTION
SEQUOIA 