Proposal
Reasoning about Actions with Large Multimodal Models

TEXAS
The University of Texas at Austin

Committee

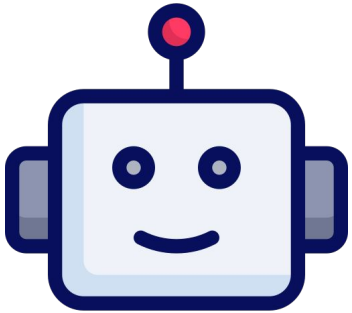
Professors Raymond Mooney (Advisor), David Harwath, Elias Stengel-Eskin, and Jacob Andreas

Vanya Cohen
PhD Student, The University of Texas at Austin

Contents

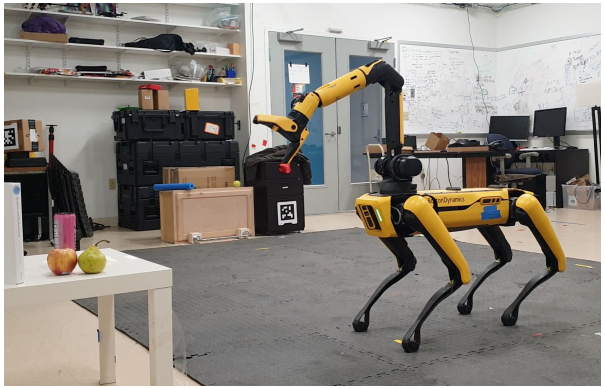
- Introduction and Background
- Completed Work
- Future Work

Contents



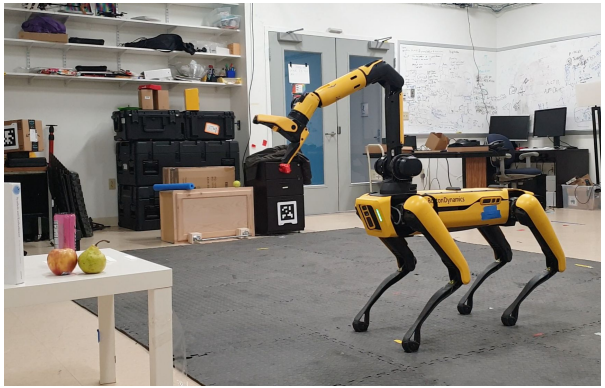
Natural language is an accessible means of specifying tasks for AI agents.

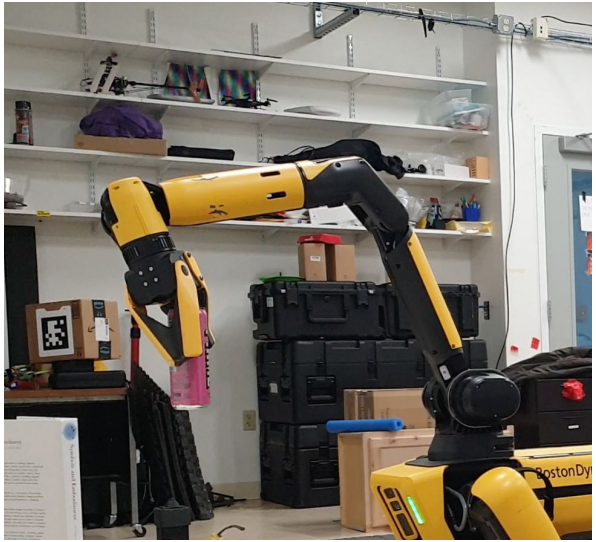
“Pick up my energy drink and place it by my running shoes.”



“Pick up my energy drink and place it by my running shoes.”

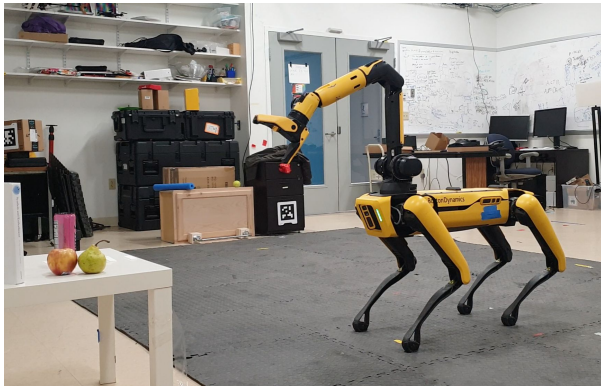
CAPE: Corrective Actions from Precondition Errors using Large Language Models
Shreyas Sundara Raman, **Vanya Cohen,** Ifrah Idrees, Eric Rosen, Ray Mooney, Stefanie Tellex, David Paulius.
ICRA 2024.

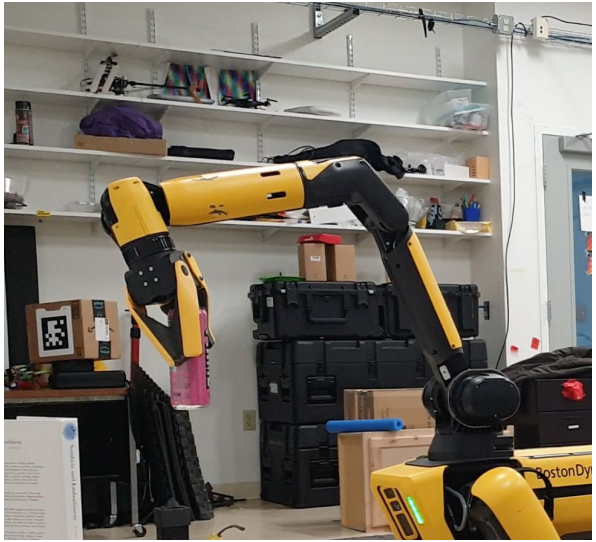




“Pick up my energy drink and place it by my running shoes.”

CAPE: Corrective Actions from Precondition Errors using Large Language Models
Shreyas Sundara Raman, **Vanya Cohen,** Ifrah Idrees, Eric Rosen, Ray Mooney, Stefanie Tellex, David Paulius.
ICRA 2024.

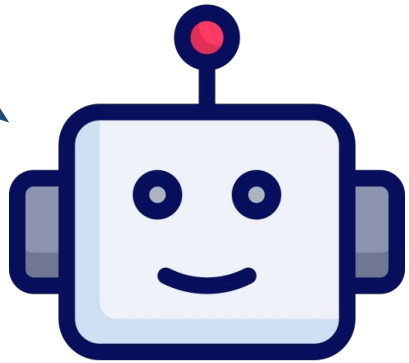


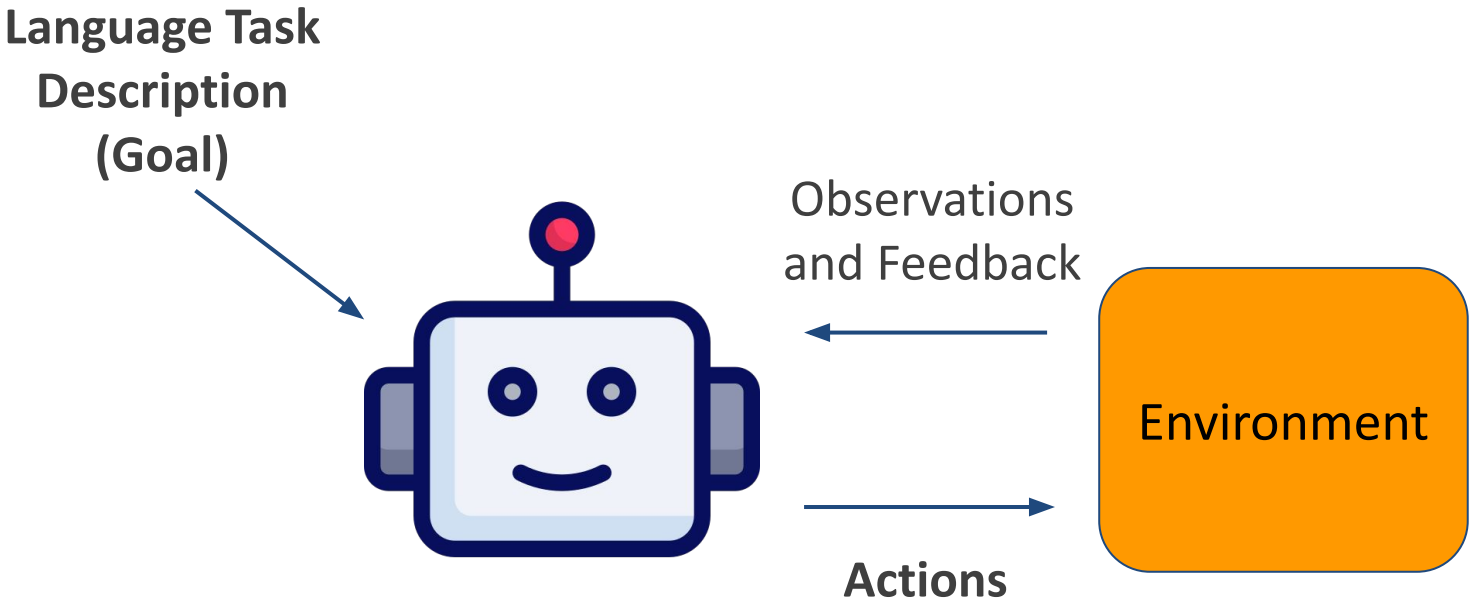


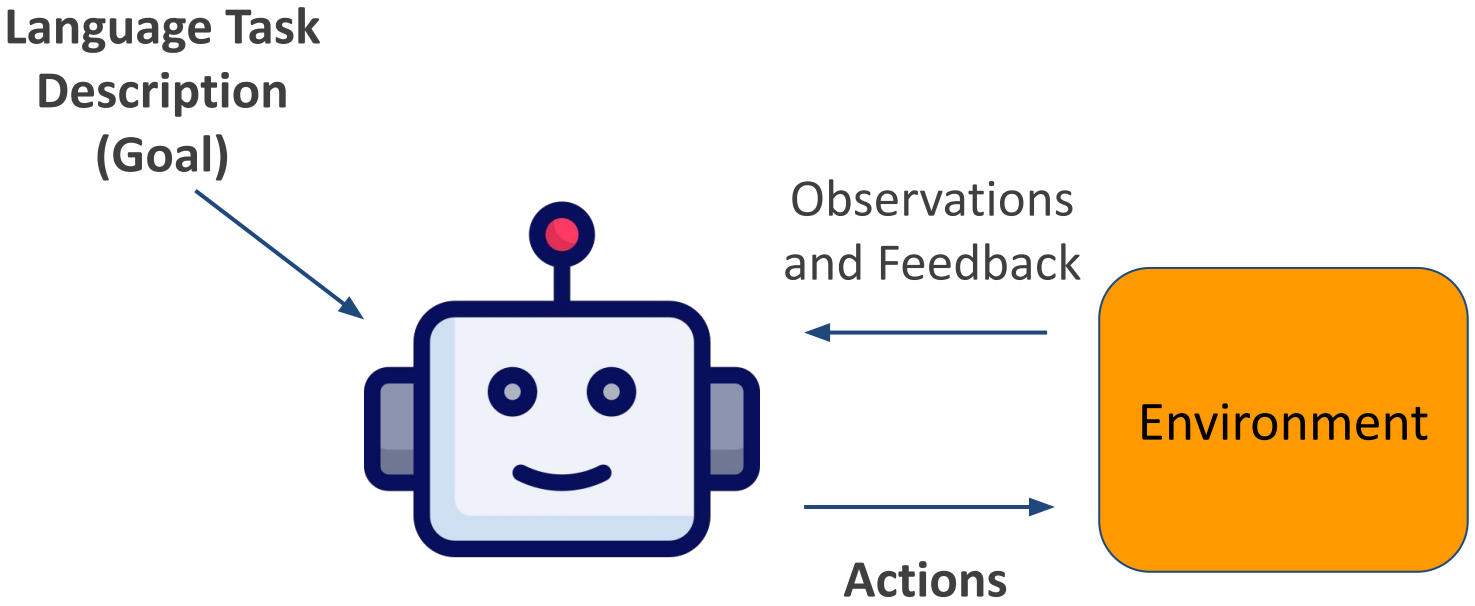
“Pick up my energy drink and place it by my running shoes.”

CAPE: Corrective Actions from Precondition Errors using Large Language Models
Shreyas Sundara Raman, **Vanya Cohen,** Ifrah Idrees, Eric Rosen, Ray Mooney, Stefanie Tellex, David Paulius.
ICRA 2024.

Language Task Description (Goal)

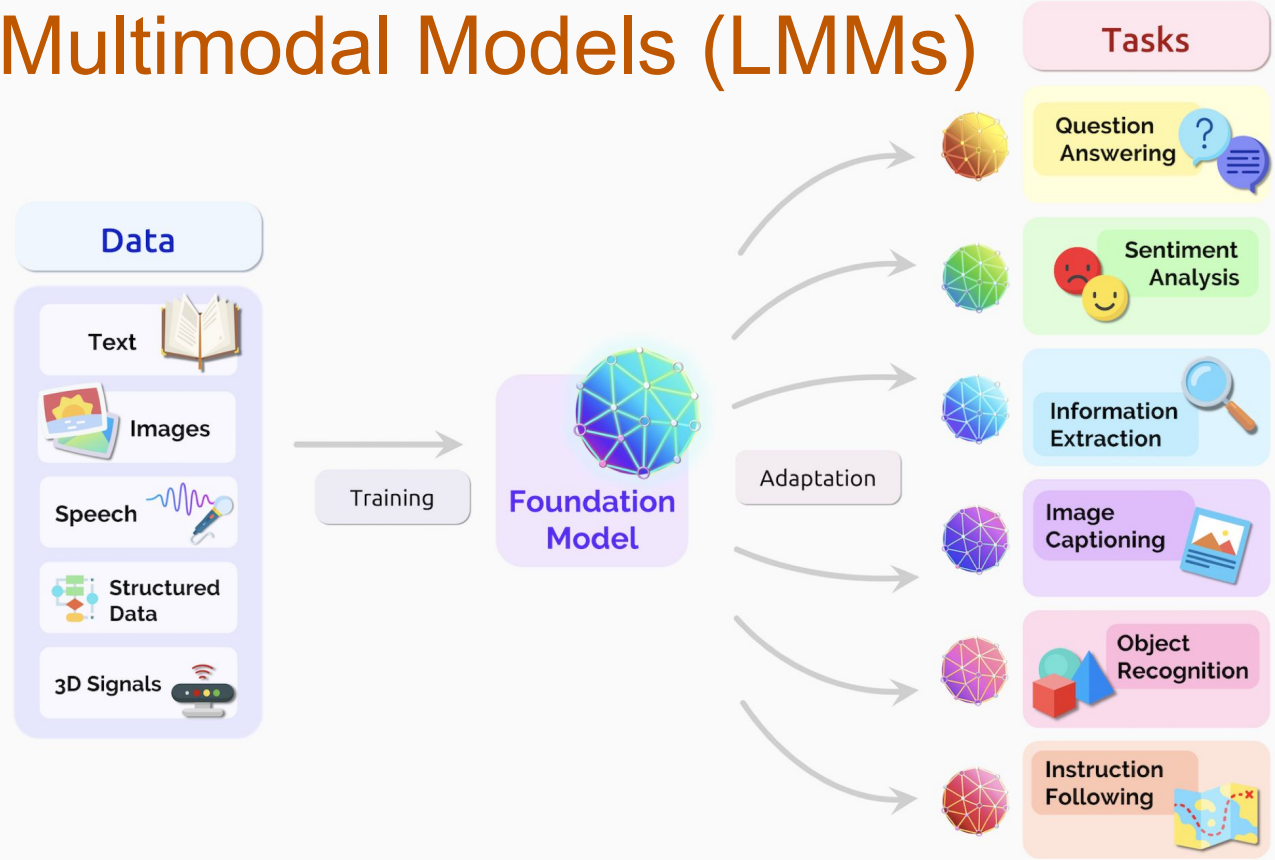



Language Task Description (Goal)
Observations and Feedback
Environment
Actions

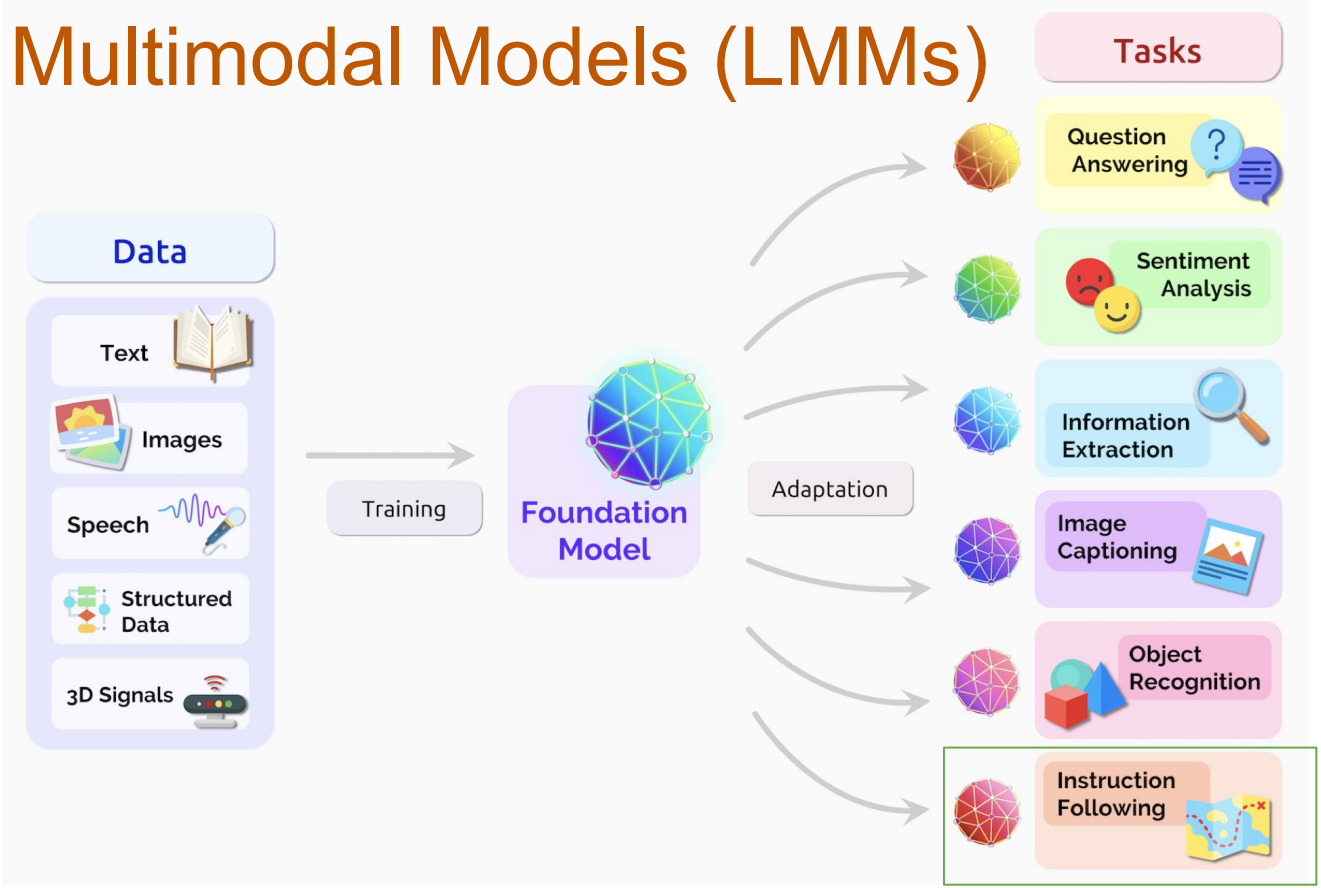


Produce **valid actions** to **accomplish a task** specified in **natural language**.

Large Multimodal Models (LMMs)



Large Multimodal Models (LMMs)



How can we evaluate and improve action reasoning in LMMs?

Action Reasoning

1. Relate **goals to action sequences.**

Action Reasoning

1. Relate **goals to action sequences.**
2. **Preconditions:** What must the world state be for the action to be afforded (executable)?

Action Reasoning

1. Relate **goals to action sequences.**
2. **Preconditions:** What must the world state be for the action to be afforded (executable)?
3. **Postconditions:** What is the world state after executing this action. I.e. World Modeling

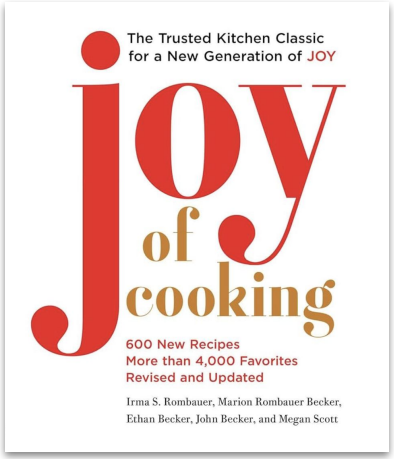
Contents

- Introduction and Related Work
- **Completed Work**
  - Using Planning to Improve Semantic Parsing of Instructional Texts (Workshop NLRSE 2023)
  - Compositional Instruction Following with Language Models and Reinforcement Learning (RLC 2025; TMLR 2024)
  - CaT-Bench: Benchmarking Language Model Understanding of Causal and Temporal Dependencies in Plans (EMNLP 2025)
- Future Work

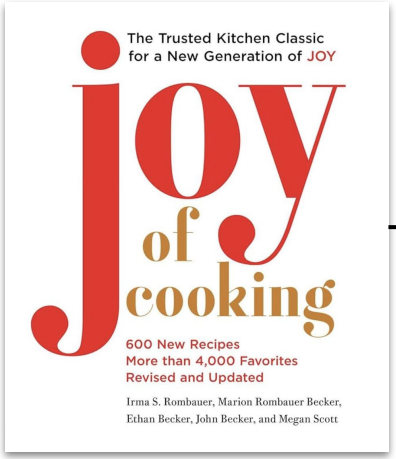
Using Planning to Improve Semantic Parsing of Instructional Texts

Vanya Cohen and Raymond Mooney

Instructional Texts



Instructional Texts



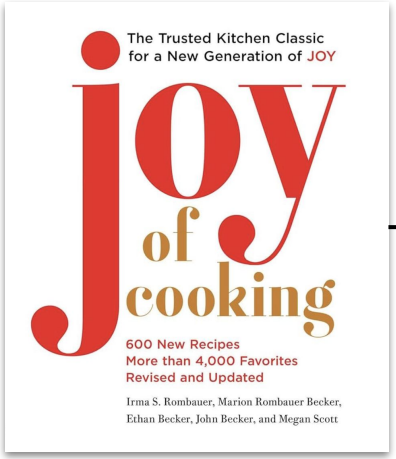
Large Pretrained Model

Executable Plan

(get-ingredient flour)
(get-ingredient sugar)
(get-ingredient butter)
(get-ingredient eggs)
(get-ingredient chips)
(preheat-oven)
(scoop-dough)
(mix-dough)
(bake-cookies)
(cool-cookies)

May not know or respect domain constraints.

Instructional Texts

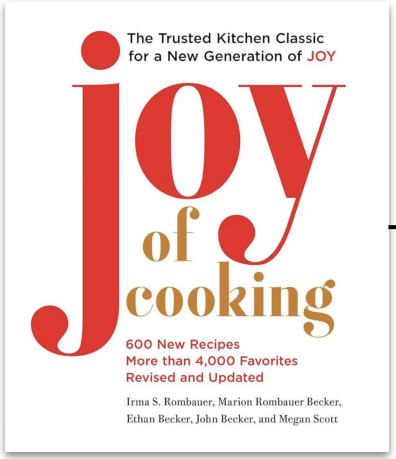


Large Pretrained Model

Executable Plan

(get-ingredient flour)
(get-ingredient sugar)
(get-ingredient butter)
(get-ingredient eggs)
(get-ingredient chips)
(preheat-oven)
(scoop-dough)
(mix-dough)
(bake-cookies)
(cool-cookies)

Instructional Texts



May not know or respect domain constraints.

Large Pretrained Model

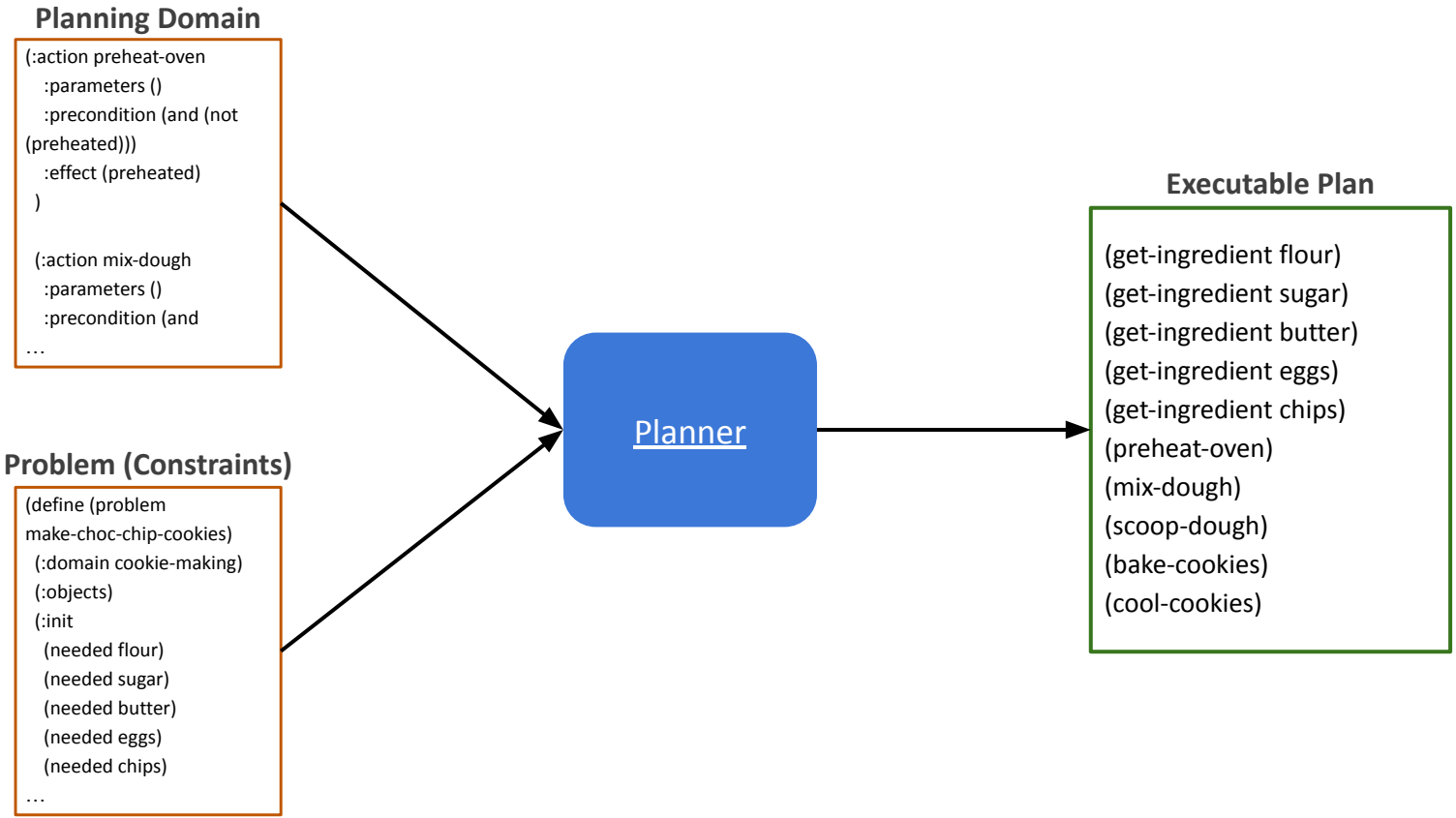
Executable Plan

(get-ingredient flour)
(get-ingredient sugar)
(get-ingredient butter)
(get-ingredient eggs)
(get-ingredient chips)
(preheat-oven)
(scoop-dough)
(mix-dough)
(bake-cookies)
(cool-cookies)

Introduction

- Few-shot semantic parsing[1] of long-form instructional texts poses unique challenges.
  - Long context dependencies.
  - Ambiguous, domain-specific language.
    - Omitted and implied steps.

1. Richard Shin, Benjamin Van Durme. Few-Shot Semantic Parsing with Language Models Trained on Code. NAACL 2022


Planning Domain
(:action preheat-oven
:parameters ()
:precondition (and (not
(preheated)))
:effect (preheated)
)
(:action mix-dough
:parameters ()
:precondition (and
…
Problem (Constraints)
(define (problem
make-choc-chip-cookies)
(:domain cookie-making)
(:objects)
(:init
(needed flour)
(needed sugar)
(needed butter)
(needed eggs)
(needed chips)
…
Planner
Executable Plan
(get-ingredient flour)
(get-ingredient sugar)
(get-ingredient butter)
(get-ingredient eggs)
(get-ingredient chips)
(preheat-oven)
(mix-dough)
(scoop-dough)
(bake-cookies)
(cool-cookies)

Method

- Utilize ***planning domain information*** to improve quality of generated **semantic parses (plans*).***
- Ensure generated plans are **executable**: each action has **satisfied preconditions**.

Method

- Planning-Augmented Semantic Parsing
  - Symbolic-planning-based decoder.
  - Ranks and corrects candidate parses.
  - Combines strength of LLMs and classical AI planning.

Datasets

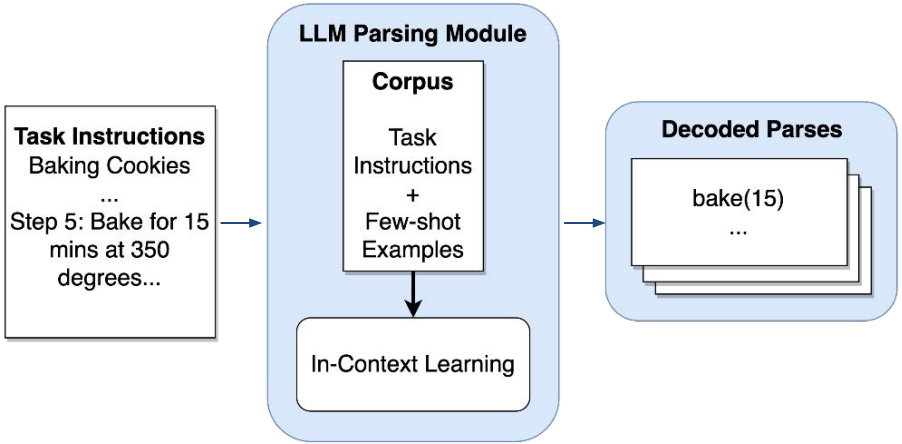
- Cooking recipes with ground-truth plans
  - Describe the steps needed to make the recipe.
  - **Bollini et al. 2013** and **Tasse and Smith 2008.**

Method

Task Instructions
Baking Cookies
...
Step 5: Bake for 15 mins at 350 degrees...

Pipeline starts with input instructions

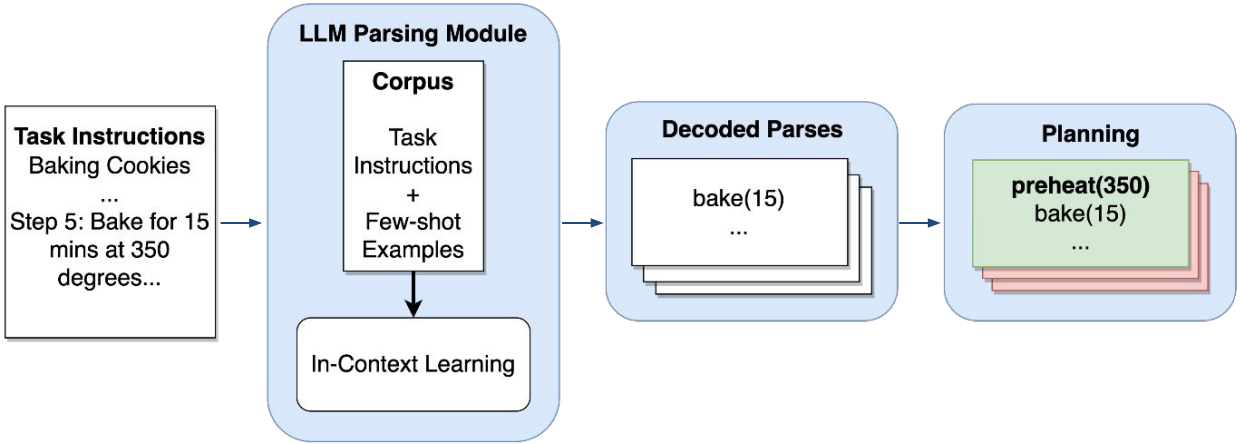
Method



Few-shot semantic parsing to translate the instructions into a sequence of operators. Produce ten candidates

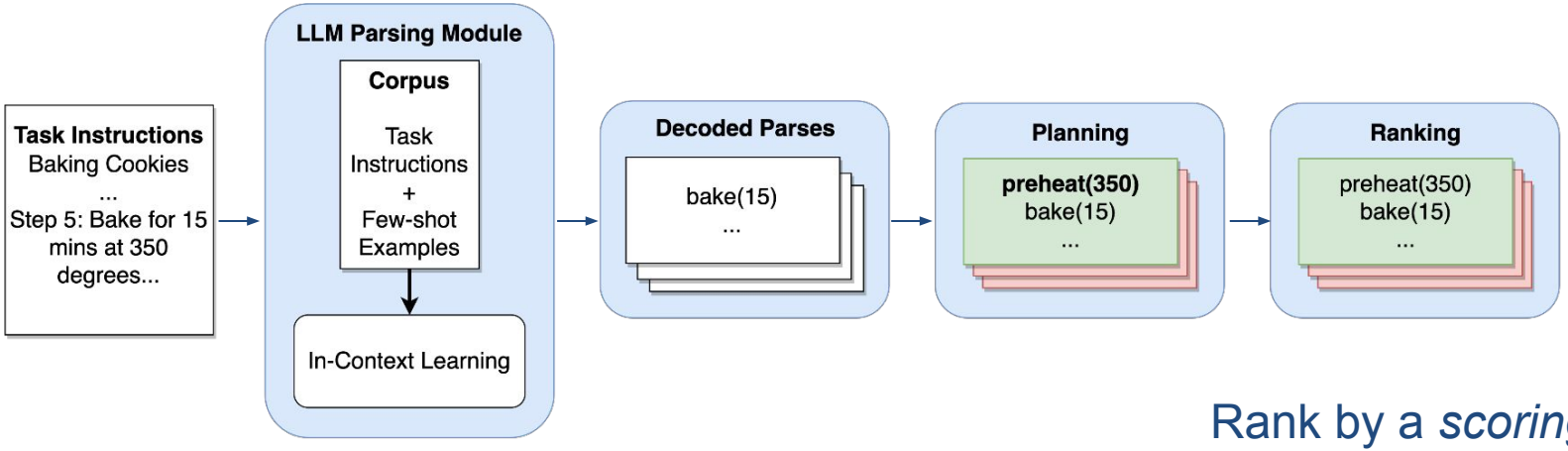
OpenAI Davinci Codex (Chen et al. 2021)

Method



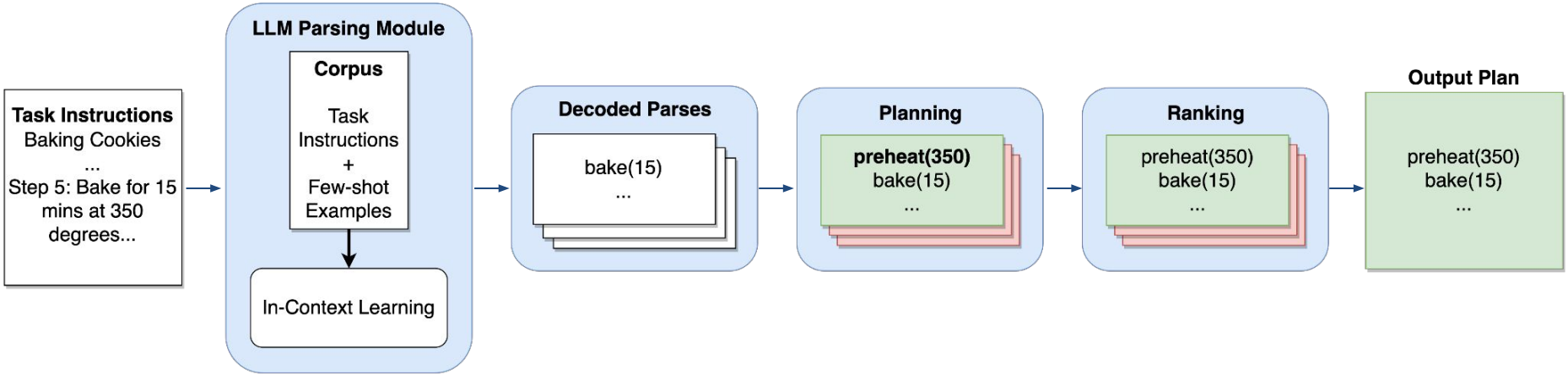
bake() preconditions not met, planning adds a preheat step

Method



Rank by a *scoring function* that prioritizes executable plans

Method


Task Instructions
Baking Cookies
...
Step 5: Bake for 15 mins at 350 degrees...
LLM Parsing Module
Corpus
Task Instructions + Few-shot Examples
In-Context Learning
Decoded Parses
bake(15)
...
Planning
preheat(350)
bake(15)
...
Ranking
preheat(350)
bake(15)
...
Output Plan
preheat(350)
bake(15)
...

Method

- Rank candidate parses (ten candidates)
  - **Minimize** syntax errors ***(SE)***, precondition errors ***(PE)***
  - **Minimize** the number of steps that need to be added to make the parsed plan executable ***(AS)***
  - **Maximize** the probability of all the plan steps **(*ln* P_t)**
- Output the highest scoring plan with added steps.

Experiments

- Evaluation Metrics
  - Longest Common Subsequence (LCS)
  - Plan Steps F1: harmonic mean of precision and recall of generated steps
  - Precondition Errors (PE) and Syntax Errors (SE)

Experiments

- Evaluation Metrics
  - Longest Common Subsequence (LCS)
  - Plan Steps F1: harmonic mean of precision and recall of generated steps
  - Precondition Errors (PE) and Syntax Errors (SE)

- Experimental Settings
  - **Rank (PPL):** Selects the plan with the lowest perplexity
  - **Rank:** Ranks the plans by the scoring function without correcting precondition errors
  - **Rank + Plan:** Our full ranking method with planning to correct errors

Results (Bollini et al. 2013)

Models	(Bollini et al., 2013) LCS↑	PE↓	SE↓	F1 ↑
Rank (PPL)				
Davinci Codex, E=1	0.897 ± 0.008	0.962 ± 0.685	0.042 ± 0.008	0.784 ± 0.004
Davinci Codex, E=5	0.949 ± 0.005	0.198 ± 0.009	0.002 ± 0.004	0.863 ± 0.003
Re-Rank				
Davinci Codex, E=1	0.901 ± 0.008	0.382 ± 0.037	0.025 ± 0.008	0.798 ± 0.002
Davinci Codex, E=5	0.952 ± 0.005	0.120 ± 0.015	0.002 ± 0.004	$\mathbf{0.868 \pm 0.002}$
Re-Rank + Plan				
Davinci Codex, E=1	0.903 ± 0.008	0.143 ± 0.033	0.025 ± 0.008	0.807 ± 0.002
Davinci Codex, E=5	0.952 ± 0.005	$\mathbf{0.033 \pm 0.000}$	0.002 ± 0.004	$\mathbf{0.870 \pm 0.002}$

Results (Bollini et al. 2013)

Models	(Bollini et al., 2013)			
	LCS↑	PE↓	SE↓	F1 ↑
Rank (PPL)				
Davinci Codex, E=1	0.897 ± 0.008	0.962 ± 0.685	0.042 ± 0.008	0.784 ± 0.004
Davinci Codex, E=5	0.949 ± 0.005	0.198 ± 0.009	0.002 ± 0.004	0.863 ± 0.003
Re-Rank				
Davinci Codex, E=1	0.901 ± 0.008	0.382 ± 0.037	0.025 ± 0.008	0.798 ± 0.002
Davinci Codex, E=5	0.952 ± 0.005	0.120 ± 0.015	0.002 ± 0.004	$\mathbf{0.868 \pm 0.002}$
Re-Rank + Plan				
Davinci Codex, E=1	0.903 ± 0.008	0.143 ± 0.033	0.025 ± 0.008	0.807 ± 0.002
Davinci Codex, E=5	0.952 ± 0.005	$\mathbf{0.033 \pm 0.000}$	0.002 ± 0.004	$\mathbf{0.870 \pm 0.002}$

Results (Bollini et al. 2013)

Models	(Bollini et al., 2013)			
	LCS↑	PE↓	SE↓	F1 ↑
Rank (PPL)				
Davinci Codex, E=1	0.897 ± 0.008	0.962 ± 0.685	0.042 ± 0.008	0.784 ± 0.004
Davinci Codex, E=5	0.949 ± 0.005	0.198 ± 0.009	0.002 ± 0.004	0.863 ± 0.003
Re-Rank				
Davinci Codex, E=1	0.901 ± 0.008	0.382 ± 0.037	0.025 ± 0.008	0.798 ± 0.002
Davinci Codex, E=5	0.952 ± 0.005	0.120 ± 0.015	0.002 ± 0.004	$\mathbf{0.868 \pm 0.002}$
Re-Rank + Plan				
Davinci Codex, E=1	0.903 ± 0.008	0.143 ± 0.033	0.025 ± 0.008	0.807 ± 0.002
Davinci Codex, E=5	0.952 ± 0.005	$\mathbf{0.033 \pm 0.000}$	0.002 ± 0.004	$\mathbf{0.870 \pm 0.002}$

Results (Tasse and Smith 2008)

Models	(Tasse and Smith, 2008)			
	LCS↑	PE↓	SE↓	F1 ↑
Rank (PPL)				
Davinci Codex, E=1	0.692 ± 0.003	0.827 ± 0.086	0.875 ± 0.199	0.443 ± 0.002
Re-Rank				
Davinci Codex, E=1	0.695 ± 0.004	0.293 ± 0.016	$\mathbf{0.226 \pm 0.024}$	0.446 ± 0.001
Re-Rank + Plan				
Davinci Codex, E=1	0.695 ± 0.003	$\mathbf{0.000 \pm 0.000}$	$\mathbf{0.237 \pm 0.018}$	0.446 ± 0.001

Results (Tasse and Smith 2008)

Models	(Tasse and Smith, 2008)			
	LCS↑	PE↓	SE↓	F1 ↑
Rank (PPL)				
Davinci Codex, E=1	0.692 ± 0.003	0.827 ± 0.086	0.875 ± 0.199	0.443 ± 0.002
Re-Rank				
Davinci Codex, E=1	0.695 ± 0.004	0.293 ± 0.016	$\mathbf{0.226 \pm 0.024}$	0.446 ± 0.001
Re-Rank + Plan				
Davinci Codex, E=1	0.695 ± 0.003	$\mathbf{0.000 \pm 0.000}$	$\mathbf{0.237 \pm 0.018}$	0.446 ± 0.001

Results (Tasse and Smith 2008)

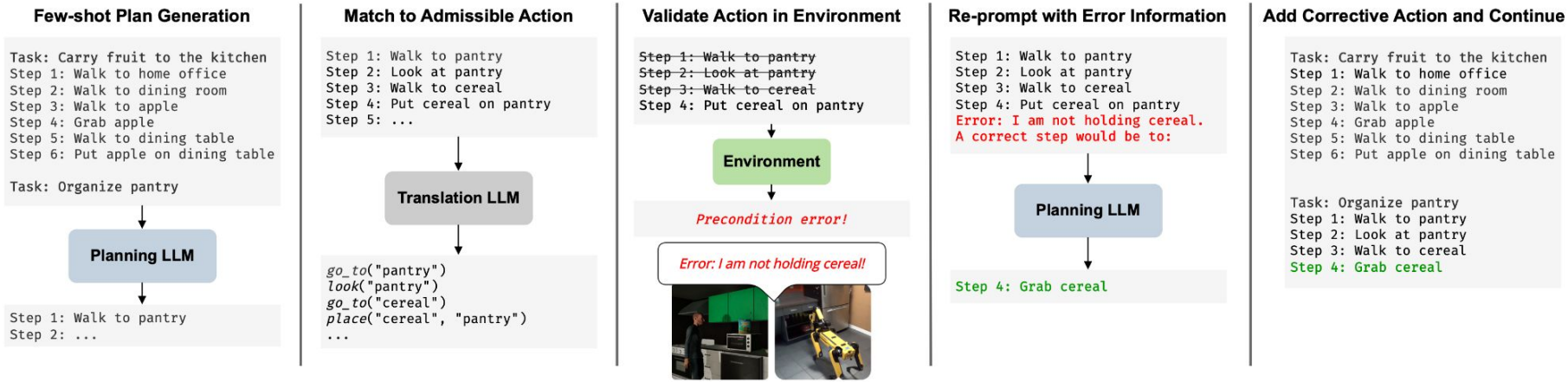
Models	(Tasse and Smith, 2008)			
	LCS↑	PE↓	SE↓	F1 ↑
Rank (PPL)				
Davinci Codex, E=1	0.692 ± 0.003	0.827 ± 0.086	0.875 ± 0.199	0.443 ± 0.002
Re-Rank				
Davinci Codex, E=1	0.695 ± 0.004	0.293 ± 0.016	$\mathbf{0.226 \pm 0.024}$	0.446 ± 0.001
Re-Rank + Plan				
Davinci Codex, E=1	0.695 ± 0.003	$\mathbf{0.000 \pm 0.000}$	$\mathbf{0.237 \pm 0.018}$	0.446 ± 0.001

Conclusion

- Neuro-symbolic approach generates semantic parses that are valid plans.
- Reduces precondition errors while maintaining content similarity to ground-truth plans.

CAPE: Corrective Actions from Precondition Errors using Large Language Models (ICRA 2024)

Shreyas Sundara Raman, **Vanya Cohen**, Ifrah Idrees, Eric Rosen, Ray Mooney, Stefanie Tellex, David Paulius.

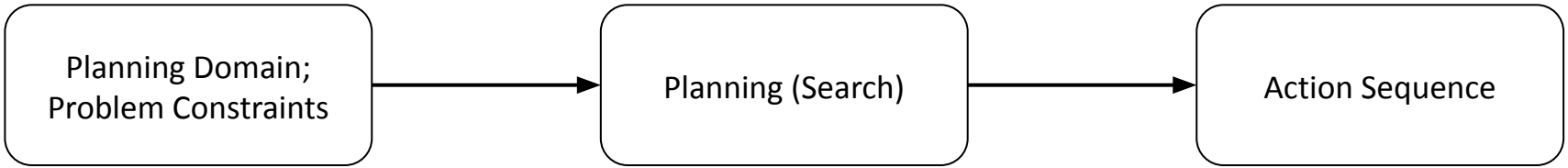


Compositional Instruction Following with Language Models and Reinforcement Learning

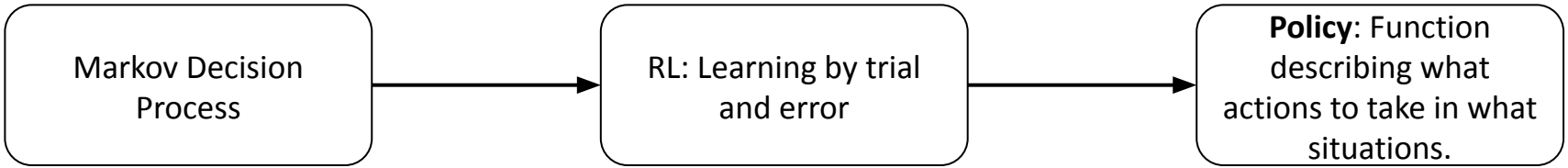
Vanya Cohen*, Geraud Nangue Tasse*, Nakul Gopalan, Steven James, Matthew Gombolay, Raymond Mooney, Benjamin Rosman

Reinforcement Learning Conference 2025; Transactions on Machine Learning Research 2024

Planning

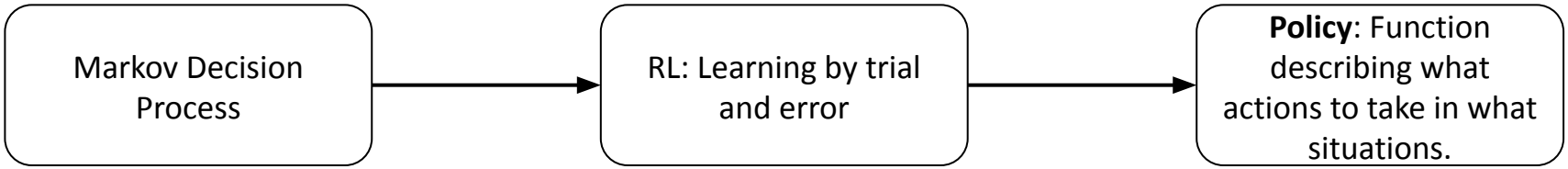

Planning Domain;
Problem Constraints
Planning (Search)
Action Sequence

Reinforcement Learning (RL)



Low sample efficiency

Reinforcement Learning (RL)



Reinforcement Learning

Domain

Markov Decision Process

States: S
Actions: A
Reward(s, a) → scalar
Transition(s, a) → s'

RL Algorithm

E.g. Policy Gradient[1]; Deep Q Networks[2]

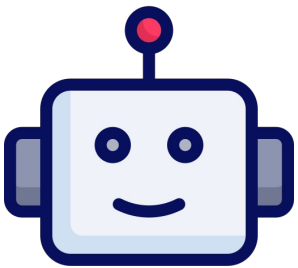
Policy Representation

Policy Function

$$\pi(a|s) = P(A_t = a \mid S_t = s)$$

Value Function

$$Q^{\pi}(s,a) = \mathbb{E}_{\pi}\left[\sum_{t=0}^{\infty} \gamma^t r_t \,\middle|\, S_0 = s, A_0 = a\right]$$



State Observation and Reward

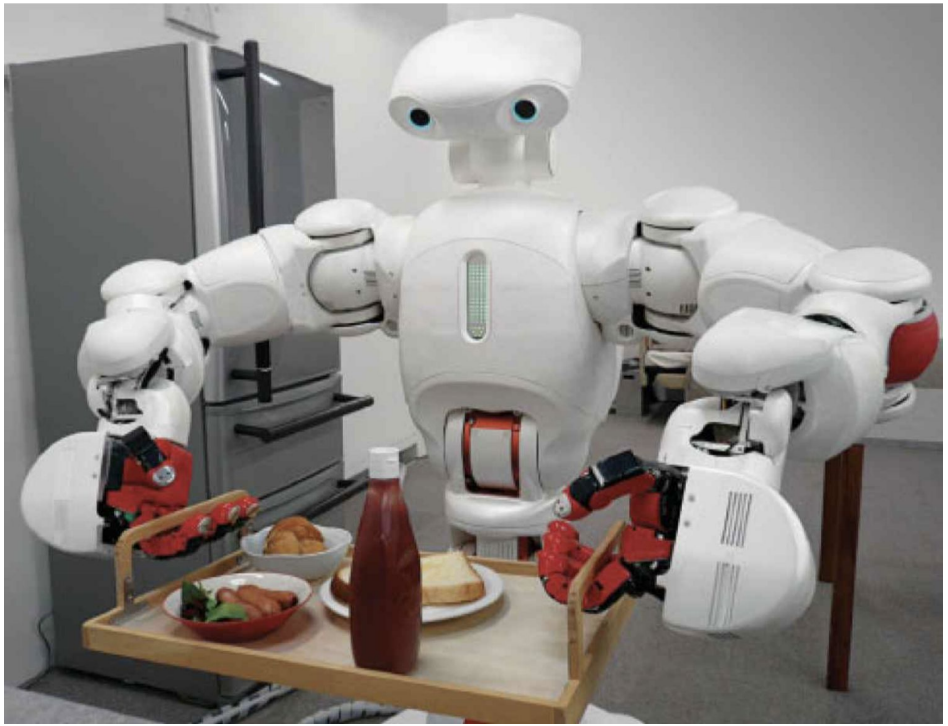
Actions

Environment

1. Richard Sutton and Andrew Barto. Reinforcement Learning: An Introduction. 2018
2. Minh et al. Playing Atari with Deep Reinforcement Learning. 2013

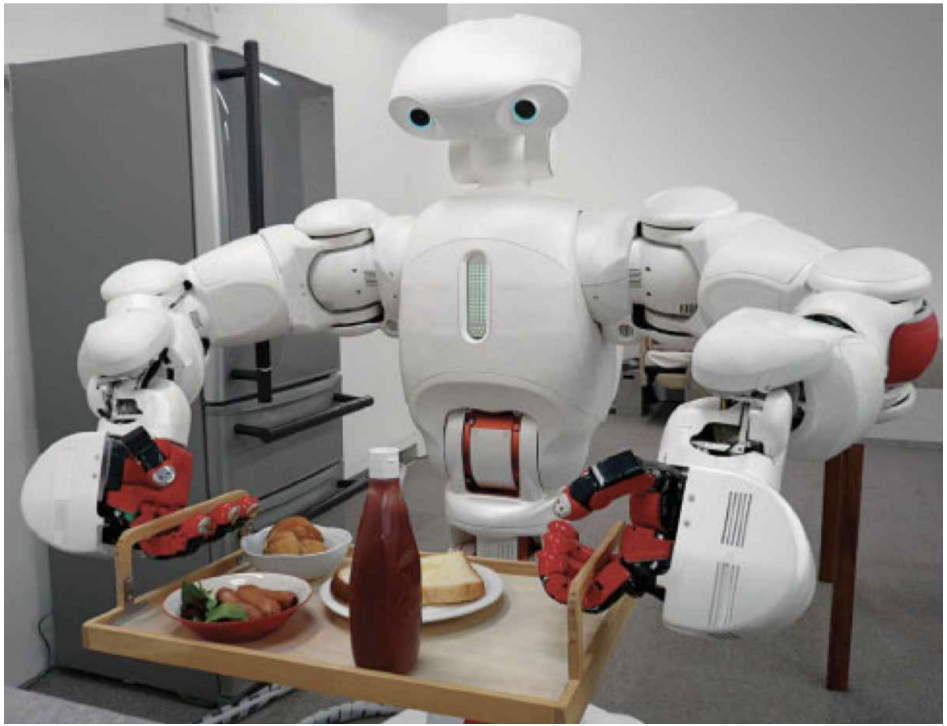
Language and RL Tasks Share Compositional Structure

- "**Serve breakfast** with **plain toast** *and* **ketchup**…"



Language and RL Tasks Share Compositional Structure

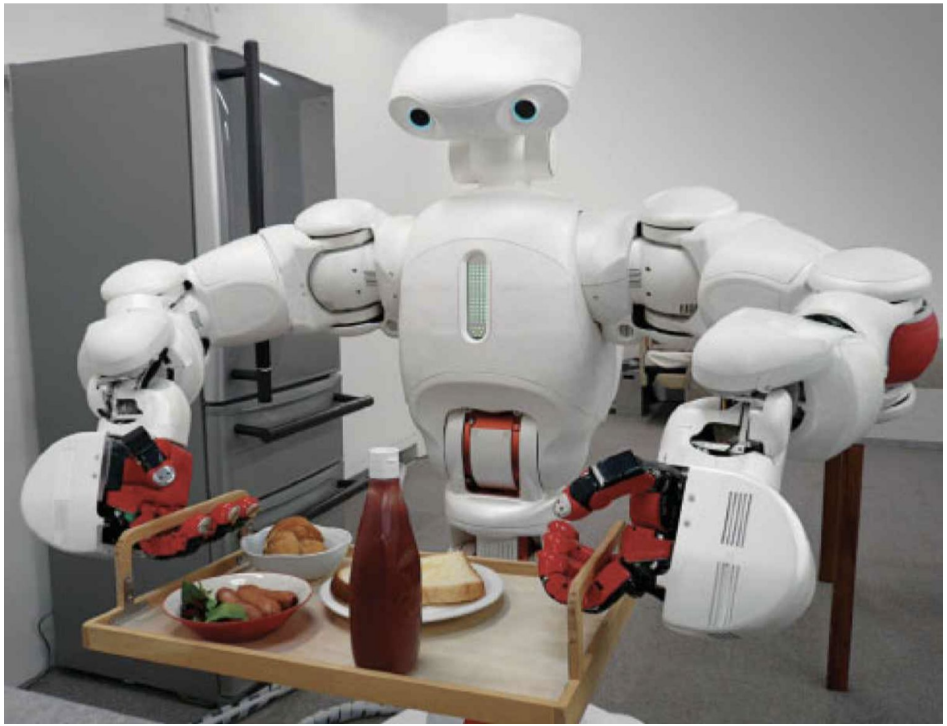
- "**Serve breakfast** with **plain toast** *and* **ketchup**…"
- Compose existing policies to perform tasks with minimal training.



1. Lake, B. M., & Baroni, M. (2018). Generalization without systematicity: On the compositional skills of sequence-to-sequence recurrent networks. Proceedings of the 35th International Conference on Machine Learning.

Language and RL Tasks Share Compositional Structure

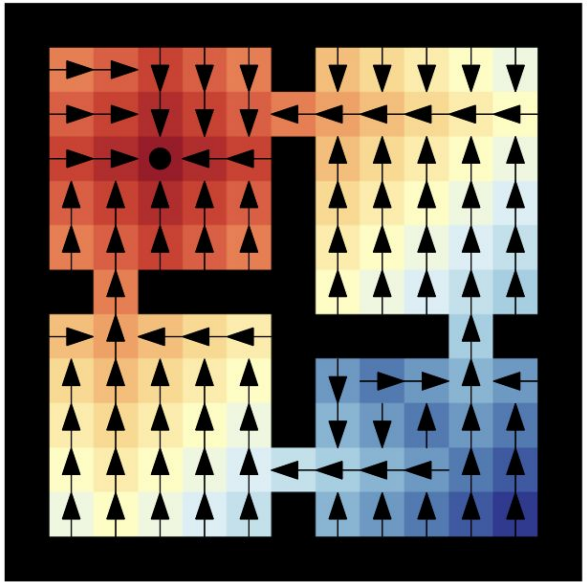
- "**Serve breakfast** with **plain toast** *and* **ketchup**…"
- Compose existing policies to perform tasks with minimal training.
- Neural networks **struggle to generalize compositionally**[1].



1. Lake, B. M., & Baroni, M. (2018). Generalization without systematicity: On the compositional skills of sequence-to-sequence recurrent networks. Proceedings of the 35th International Conference on Machine Learning.

State-Action Value Function

$$Q_\pi(s,a) = \mathbb{E}_s^\pi\left[\sum_{t=0}^{\infty} r(s_t,a_t)\right]$$

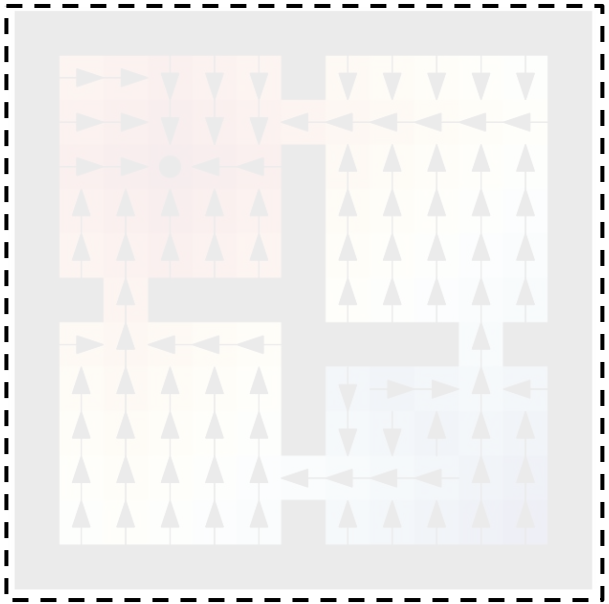


"Go to the top left room."

1. Nangue Tasse, G., James, S., & Rosman, B. (2020). A Boolean task algebra for reinforcement learning. Advances in Neural Information Processing Systems, 33, 17279–17290.
2. Nangue Tasse, G., James, S., & Rosman, B. (2022, June). World value functions: Knowledge representation for multitask reinforcement learning. Paper presented at the 5th Multi-disciplinary Conference on Reinforcement Learning and Decision Making (RLDM).

State-Action Value Function

$$Q_\pi(s,a) = \mathbb{E}_s^\pi \left[\sum_{t=0}^{\infty} r(s_t, a_t) \right]$$



“Go to the top right room.”

1. Nangue Tasse, G., James, S., & Rosman, B. (2020). A Boolean task algebra for reinforcement learning. Advances in Neural Information Processing Systems, 33, 17279–17290.
2. Nangue Tasse, G., James, S., & Rosman, B. (2022, June). World value functions: Knowledge representation for multitask reinforcement learning. Paper presented at the 5th Multi-disciplinary Conference on Reinforcement Learning and Decision Making (RLDM).

State-Action Value Function

$$Q_\pi(s,a) = \mathbb{E}_s^\pi \left[\sum_{t=0}^{\infty} r(s_t, a_t) \right]$$

Learn a new value function from scratch.

“Go to the top right room.”

1. Nangue Tasse, G., James, S., & Rosman, B. (2020). A Boolean task algebra for reinforcement learning. Advances in Neural Information Processing Systems, 33, 17279–17290.
2. Nangue Tasse, G., James, S., & Rosman, B. (2022, June). World value functions: Knowledge representation for multitask reinforcement learning. Paper presented at the 5th Multi-disciplinary Conference on Reinforcement Learning and Decision Making (RLDM).

World Value Functions (Tasse et al. 2020, 2022)

- Add a goal *g* to the Q function.
- WVF represents how to achieve all goals **and** their values.
- Learn one WVF for each task in the environment we wish to compose.

$$Q_\pi(s, g, a) = \mathbb{E}_s^\pi \left[\sum_{t=0}^{\infty} \bar{r}(s_t, g, a_t) \right]$$

1. Nangue Tasse, G., James, S., & Rosman, B. (2020). A Boolean task algebra for reinforcement learning. Advances in Neural Information Processing Systems, 33, 17279–17290.
2. Nangue Tasse, G., James, S., & Rosman, B. (2022, June). World value functions: Knowledge representation for multitask reinforcement learning. Paper presented at the 5th Multi-disciplinary Conference on Reinforcement Learning and Decision Making (RLDM).

World Value Functions (Tasse et al. 2020, 2022)

- Add a goal *g* to the Q function.
- WVF represents how to achieve all goals **and** their values.
- Learn one WVF for each task in the environment we wish to compose.
- Train by **penalizing** the agent for entering a terminal state for another goal.

$$Q_\pi(s, g, a) = \mathbb{E}_s^\pi \left[\sum_{t=0}^{\infty} \bar{r}(s_t, g, a_t) \right]$$

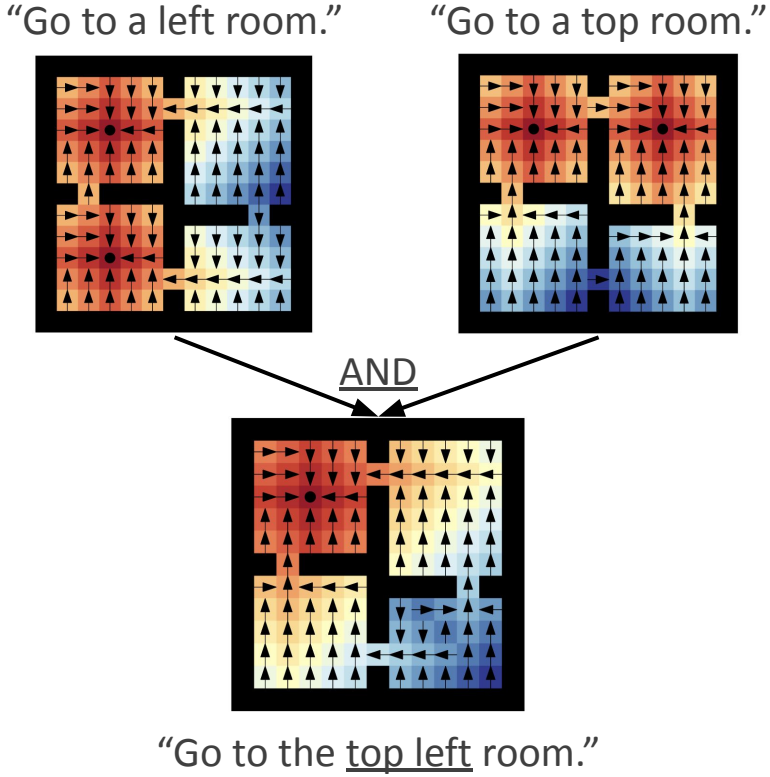
$$\bar{r}(s, g, a) = \begin{cases} \bar{r}_{MIN} & if\ g \neq s \in G \\ r(s, a) & otherwise \end{cases}$$

1. Nangue Tasse, G., James, S., & Rosman, B. (2020). A Boolean task algebra for reinforcement learning. Advances in Neural Information Processing Systems, 33, 17279–17290.
2. Nangue Tasse, G., James, S., & Rosman, B. (2022, June). World value functions: Knowledge representation for multitask reinforcement learning. Paper presented at the 5th Multi-disciplinary Conference on Reinforcement Learning and Decision Making (RLDM).

World Value Functions (Tasse et al. 2020, 2022)

For AND (conjunction) the composed WVF is given by:

$$\bar{Q}_1^* \wedge \bar{Q}_2^* : \mathcal{S} \times \mathcal{G} \times \mathcal{A} \to \mathbb{R}$$
$$(s, g, a) \mapsto \min\{\bar{Q}_1^*(s, g, a), \bar{Q}_2^*(s, g, a)\}$$

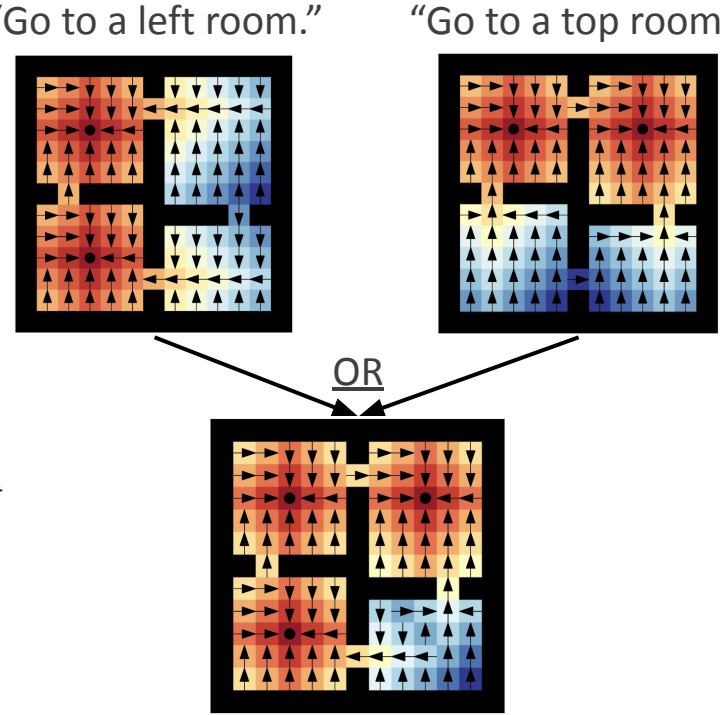


1. Nangue Tasse, G., James, S., & Rosman, B. (2020). A Boolean task algebra for reinforcement learning. Advances in Neural Information Processing Systems, 33, 17279–17290.
2. Nangue Tasse, G., James, S., & Rosman, B. (2022, June). World value functions: Knowledge representation for multitask reinforcement learning. Paper presented at the 5th Multi-disciplinary Conference on Reinforcement Learning and Decision Making (RLDM).

World Value Functions (Tasse et al. 2020, 2022)

For OR (disjunction) the composed WVF is given by:

$$\bar{Q}_1^* \vee \bar{Q}_2^* : \mathcal{S} \times \mathcal{G} \times \mathcal{A} \to \mathbb{R}$$
$$(s, g, a) \mapsto \max\{\bar{Q}_1^*(s, g, a), \bar{Q}_2^*(s, g, a)\}$$



1. Nangue Tasse, G., James, S., & Rosman, B. (2020). A Boolean task algebra for reinforcement learning. Advances in Neural Information Processing Systems, 33, 17279–17290.
2. Nangue Tasse, G., James, S., & Rosman, B. (2022, June). World value functions: Knowledge representation for multitask reinforcement learning. Paper presented at the 5th Multi-disciplinary Conference on Reinforcement Learning and Decision Making (RLDM).

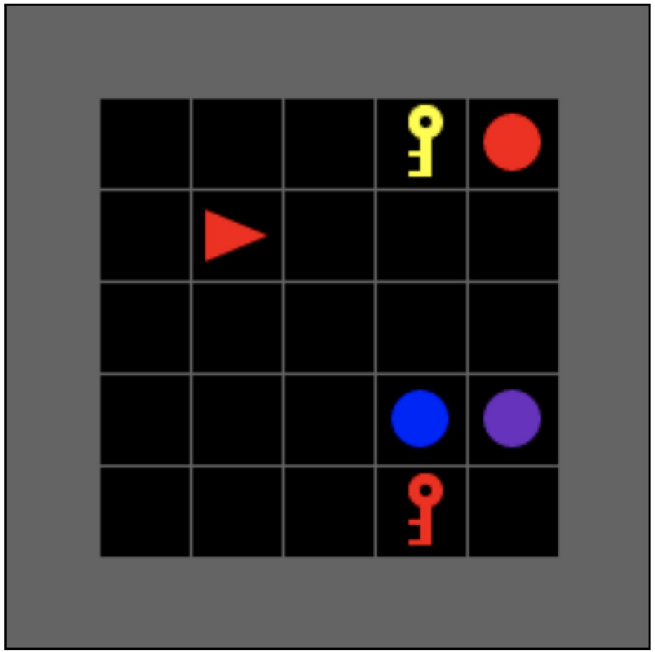
World Value Functions (WVF) (Tasse et al. 2020, 2022)

Composing WVFs

- Arbitrary expressions of **AND**, **OR**, and **NOT**.
- Can now solve a **combinatorial number** of goal reaching tasks

BabyAI (Chevalier-Boisvert et al. 2019)

- Gridworld domain consisting of language instruction tasks.

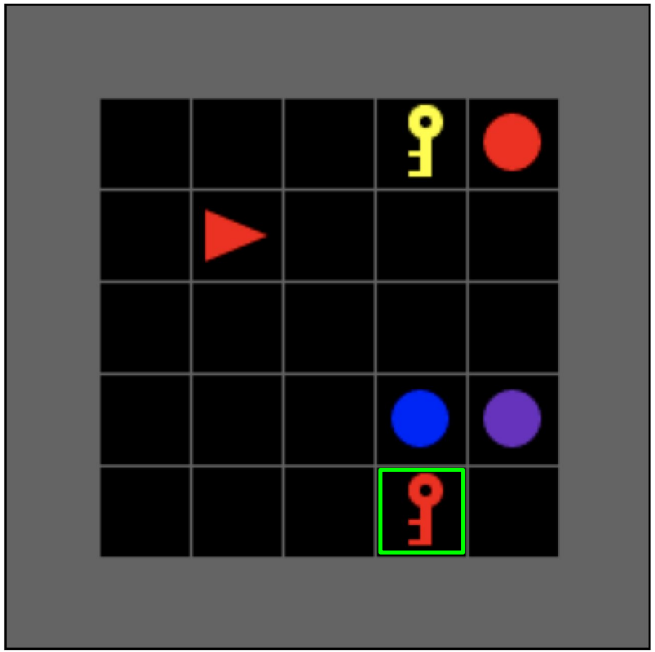


BabyAI (Chevalier-Boisvert et al. 2019)

- Gridworld domain consisting of language instruction tasks.

"Pick up a red object" ∧ "Pick up a key"

¬"Pick up a blue object" ∨ "Pick up a ball"



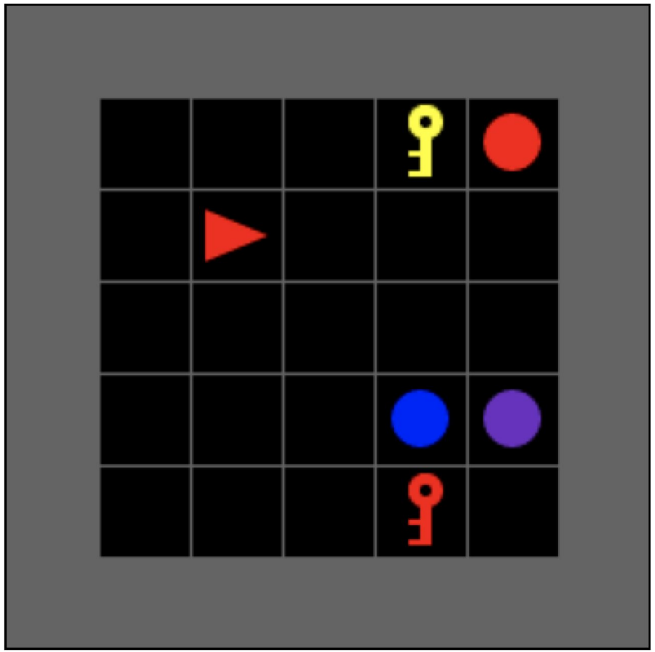
BabyAI (Chevalier-Boisvert et al. 2019)

- Gridworld domain consisting of language instruction tasks.

"Pick up a red object" ∧ "Pick up a key"

¬"Pick up a blue object" ∨ "Pick up a ball"

- Modified task set to include 162 goal reaching tasks that can be solved through **AND**, **OR**, and **NOT** expressions over object attributes.



Compositionally-Enabled RL and Language Agent (CERLLA)

- How can we use compositionality of language + value functions to generalize better?

Compositionally-Enabled RL and Language Agent (CERLLA)

- How can we use compositionality of language + value functions to generalize better?
- Must learn mapping from natural language to WVF composition.

Compositionally-Enabled RL and Language Agent (CERLLA)

- How can we use compositionality of language + value functions to generalize better?
- Must learn mapping from natural language to WVF composition.
- Idea: Use language models to translate instructions into formal language / boolean symbols (e.g. semantic parsing).

Compositionally-Enabled RL and Language Agent (CERLLA)

- But these symbols are arbitrary (just an index over WVFs) - how do we know translation is correct?

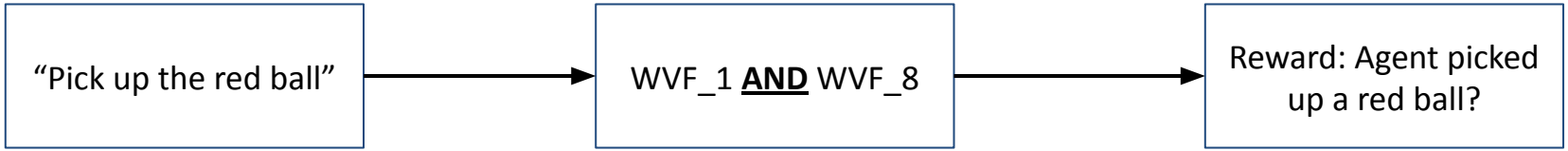
Compositionally-Enabled RL and Language Agent (CERLLA)

- But these symbols are arbitrary (just an index over WVFs) - how do we know translation is correct?
- Idea: Use environment feedback to learn the translation!

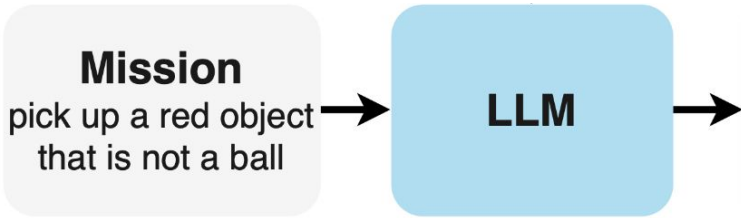
Compositionally-Enabled RL and Language Agent (CERLLA)

Core challenge: CERLLA learns to parse input commands to **arbitrary symbols** representing WVFs with **unknown semantics**, using **environment rollouts**, a much noisier form of supervision than is typical for weakly supervised parsing methods.

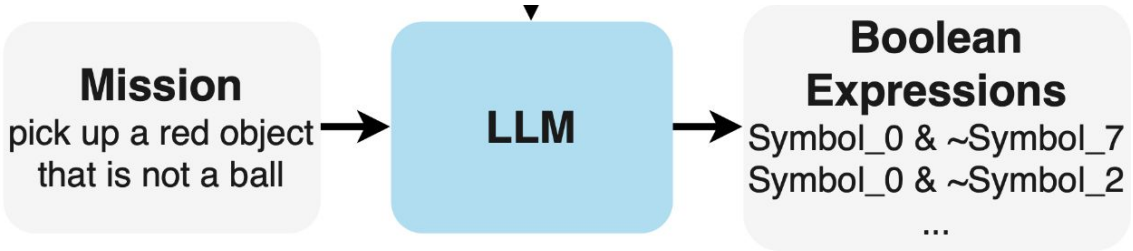
Compositionally-Enabled RL and Language Agent (CERLLA)



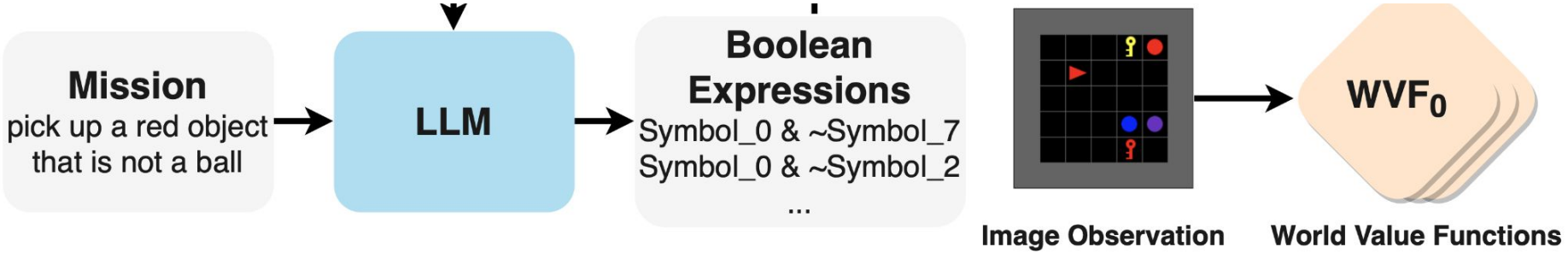
Compositionally-Enabled RL and Language Agent (CERLLA)



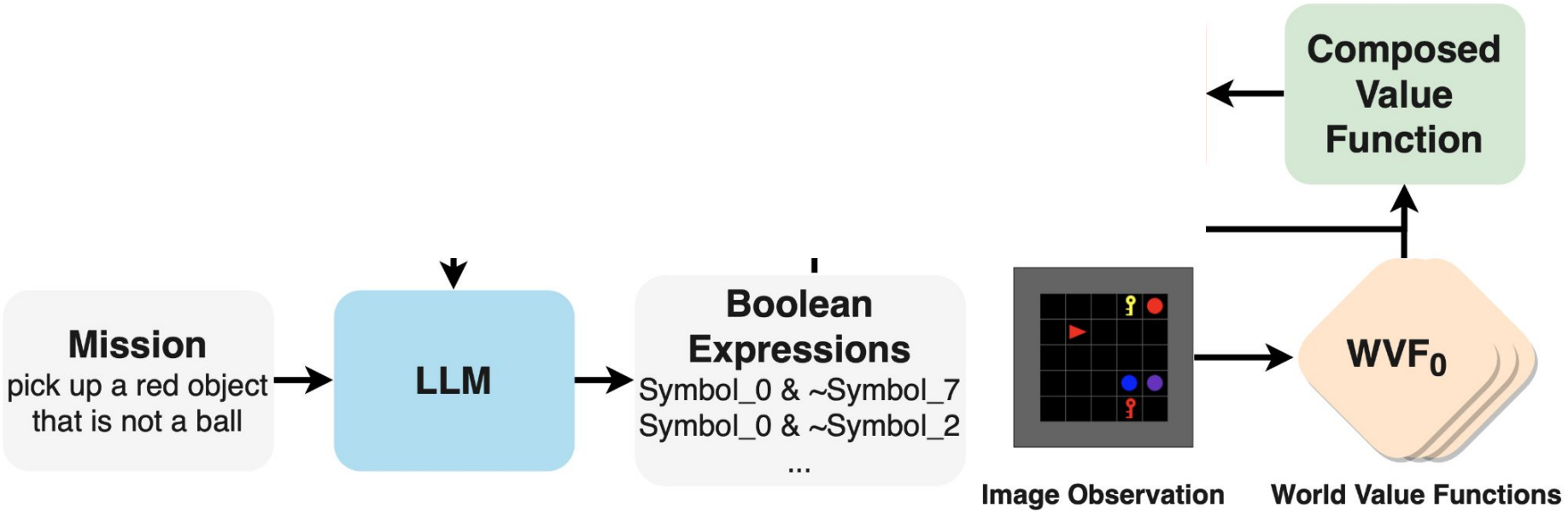
Compositionally-Enabled RL and Language Agent (CERLLA)



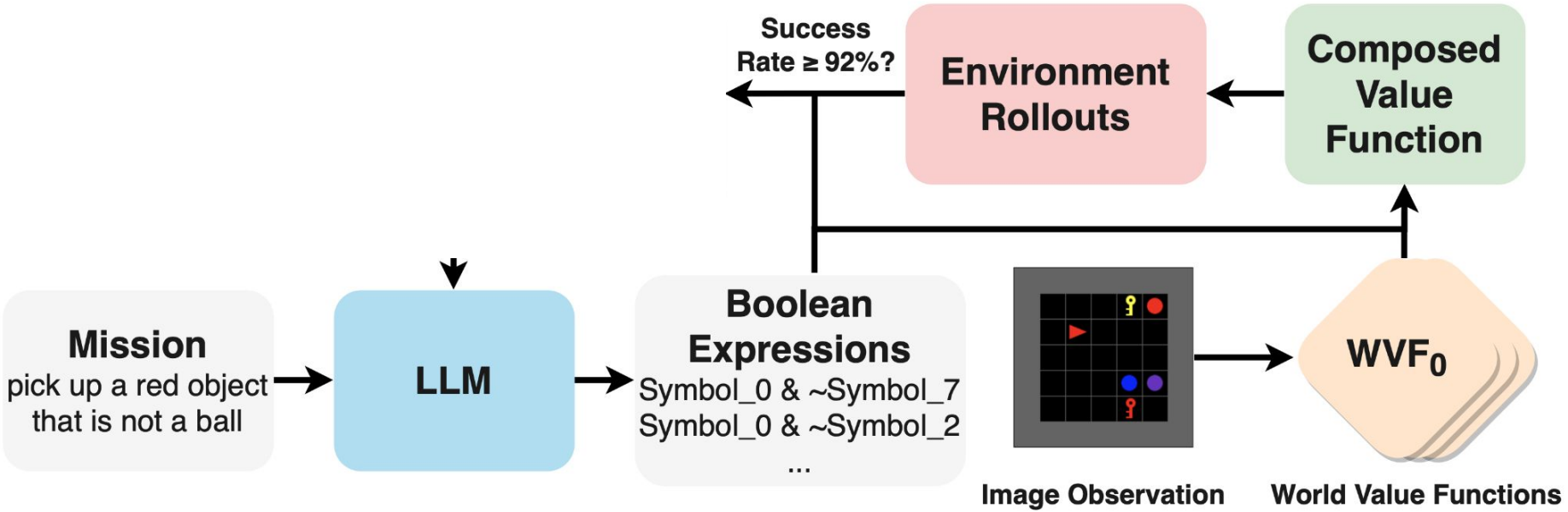
Compositionally-Enabled RL and Language Agent (CERLLA)



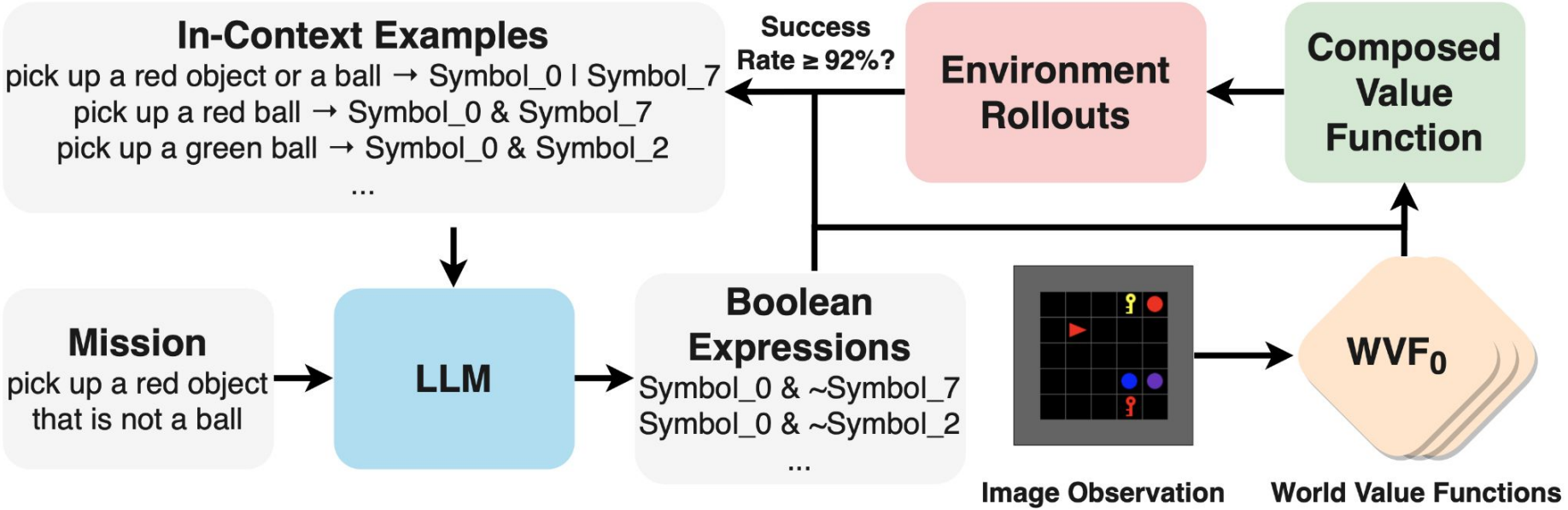
Compositionally-Enabled RL and Language Agent (CERLLA)



Compositionally-Enabled RL and Language Agent (CERLLA)



Compositionally-Enabled RL and Language Agent (CERLLA)



Experiments

- 162 tasks, learned simultaneously from vision and language.

Experiments

- 162 tasks, learned simultaneously from vision and language.
- Evaluate **sample efficiency**, and **generalization**, comparing:

Experiments

- 162 tasks, learned simultaneously from vision and language.
- Evaluate **sample efficiency**, and **generalization**, comparing:
  - CERLLA (Ours): using OpenAI's GPT-4 LM
    - CERLLA GPT-3.5

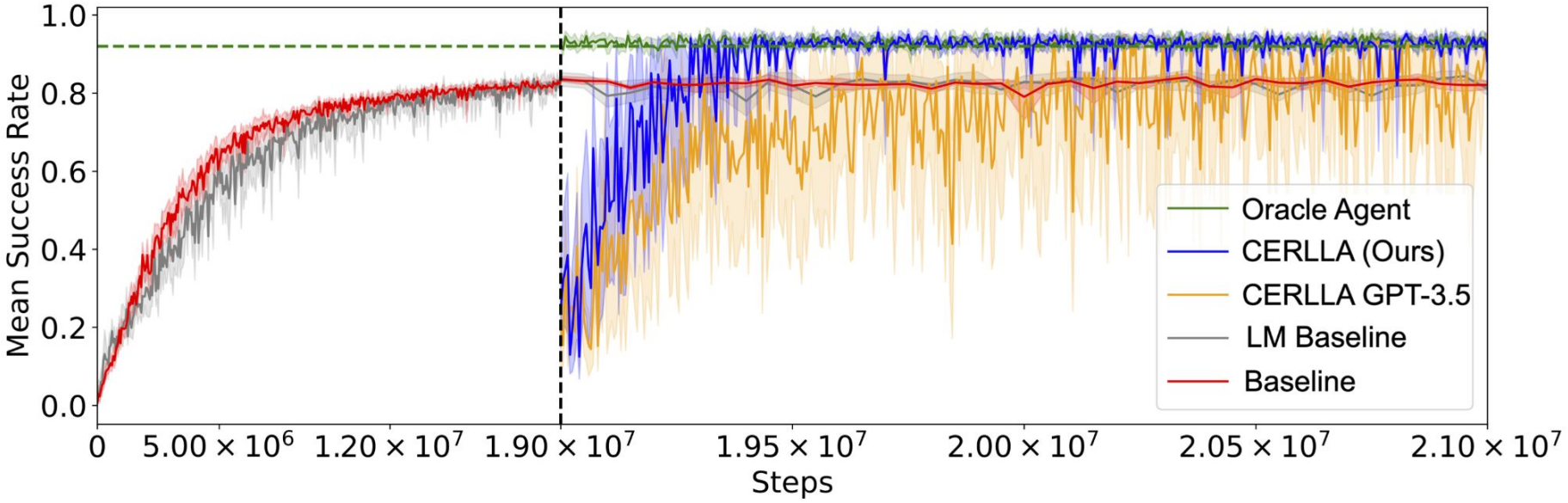
Experiments

- 162 tasks, learned simultaneously from vision and language.
- Evaluate **sample efficiency**, and **generalization**, comparing:
  - CERLLA (Ours): using OpenAI's GPT-4 LM
    - CERLLA GPT-3.5
  - Two non-compositional baseline DQNs
    - Baseline: RNN + CNN
    - LM Baseline: pretrained sentence embedding language model + CNN

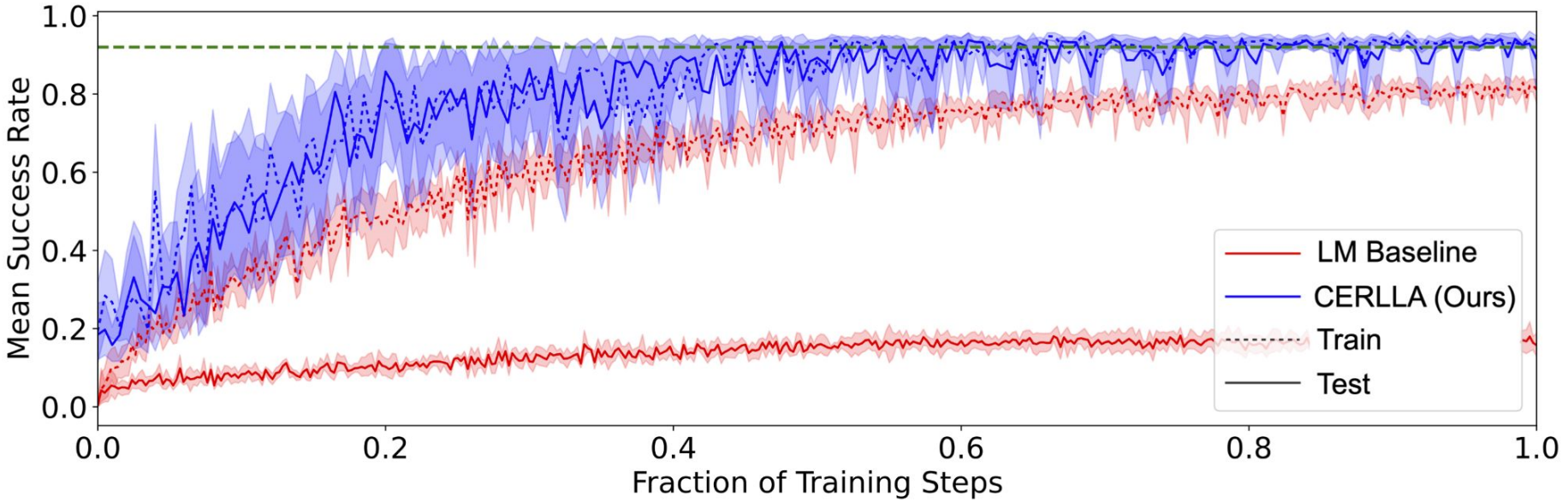
Experiments

- 162 tasks, learned simultaneously from vision and language.
- Evaluate **sample efficiency**, and **generalization**, comparing:
  - CERLLA (Ours): using OpenAI's GPT-4 LM
    - CERLLA GPT-3.5
  - Two non-compositional baseline DQNs
    - Baseline: RNN + CNN
    - LM Baseline: pretrained sentence embedding language model + CNN
  - Oracle Agent with access to the ground-truth compositional expressions for each task.

Sample Efficiency



Generalization



Conclusion

- Introduces CERLLA, a **novel semantic parsing** method based on **in-context learning** that learns from **environment feedback**.

Conclusion

- Introduces CERLLA, a **novel semantic parsing** method based on **in-context learning** that learns from **environment feedback**.
- Simultaneously learns and solves a large collection of **162 compositional vision-language-RL tasks**.

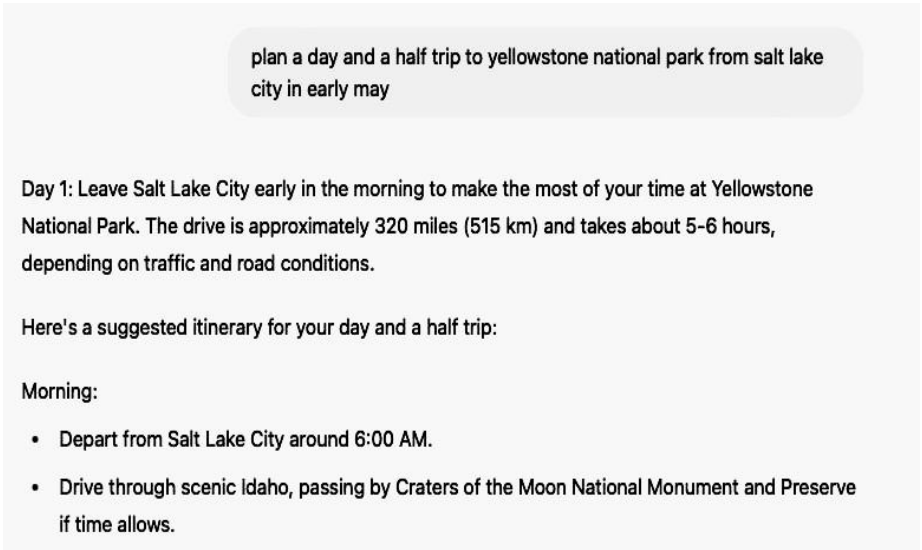
Conclusion

- Introduces CERLLA, a **novel semantic parsing** method based on **in-context learning** that learns from **environment feedback**.
- Simultaneously learns and solves a large collection of **162 compositional vision-language-RL tasks**.
- Outperforms non-compositional baselines with respect to sample efficiency and generalization to held-out tasks.

CaT-Bench: Benchmarking Language Model Understanding of Causal and Temporal Dependencies in Plans

Yash Kumar Lal*, **Vanya Cohen***, Nathanael Chambers, Niranjan Balasubramanian, Raymond Mooney

LLMs Generate Fluent Plans

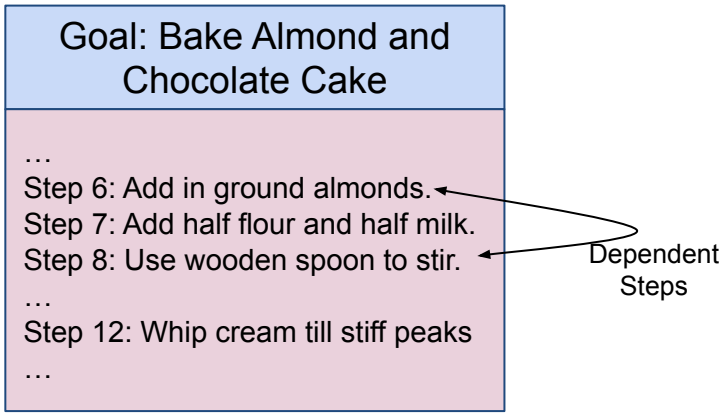


But can they reason about preconditions in plans?

Plans Contain Dependencies

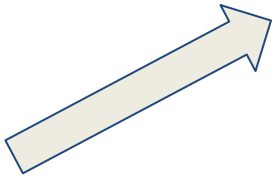
Goal: Bake Almond and Chocolate Cake
… Step 6: Add in ground almonds. Step 7: Add half flour and half milk. Step 8: Use wooden spoon to stir. … Step 12: Whip cream till stiff peaks …

Plans Contain Dependencies



Goal: Bake Almond and Chocolate Cake

…
Step 6: Add in ground almonds.
Step 7: Add half flour and half milk.
Step 8: Use wooden spoon to stir.
…
Step 12: Whip cream till stiff peaks
…



Q: Must Step 6 happen before Step 8?

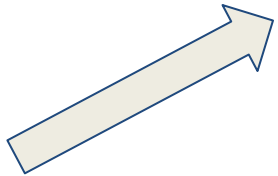
Preconditions

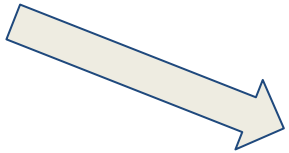
A: Yes, all ingredients have to be in bowl before stirring

Questions about dependent steps

Goal: Bake Almond and Chocolate Cake

…
Step 6: Add in ground almonds.
Step 7: Add half flour and half milk.
Step 8: Use wooden spoon to stir.
…
Step 12: Whip cream till stiff peaks
…





Q: Must Step 6 happen before Step 8?

Preconditions

A: Yes, all ingredients have to be in bowl before stirring

Questions about dependent steps

Q: Must Step 7 happen after Step 6?

Parallel Steps

A: No, almonds can be added after flour and milk

Questions about non-dependent steps

Creating 🐱 CaT-Bench

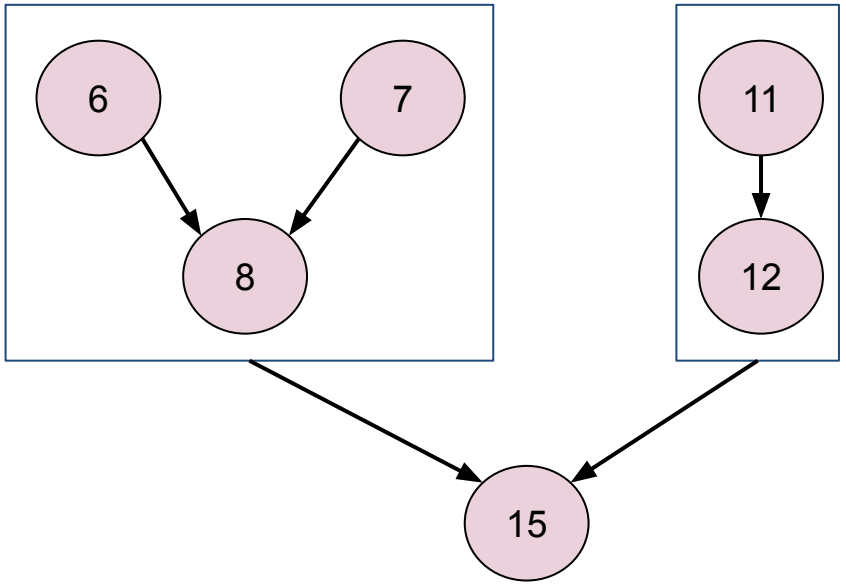
Goal: Bake Almond and Chocolate Cake
… Step 6: Add in ground almonds. Step 7: Add half flour and half milk. Step 8: Use wooden spoon to stir. … Step 12: Whip cream till stiff peaks …

Recipe Flow Graph Corpus (Yamakata et al. 2020)

Creating 🐱 CaT-Bench

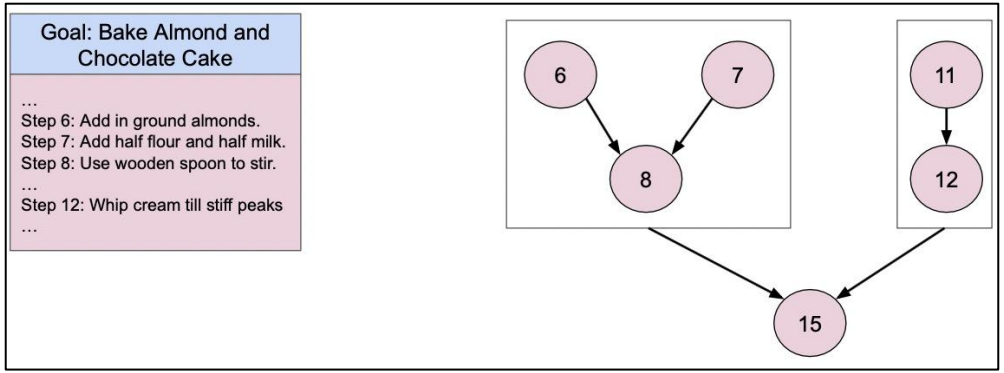
Goal: Bake Almond and Chocolate Cake

…
Step 6: Add in ground almonds.
Step 7: Add half flour and half milk.
Step 8: Use wooden spoon to stir.
…
Step 12: Whip cream till stiff peaks
…



Recipe Flow Graph Corpus (Yamakata et al. 2020)

Creating 🐱 CaT-Bench



710	Q: Must Step 6 happen before Step 8?	
710	Q: Must Step 8 happen after Step 6?	
1420	Questions about dependent steps	

	Q: Must Step 12 happen after Step 6?	710
	Q: Must Step 6 happen before Step 12?	710
	Questions about non-dependent steps	1420

Benchmarking Models

- GPT-3.5
- GPT-4-Turbo
- GPT-4o
- Llama-3-8B
- Claude-3.5-Sonnet
- Gemini-1.0-Pro
- Gemini-1.5-Pro
- Gemini-1.5-Flash

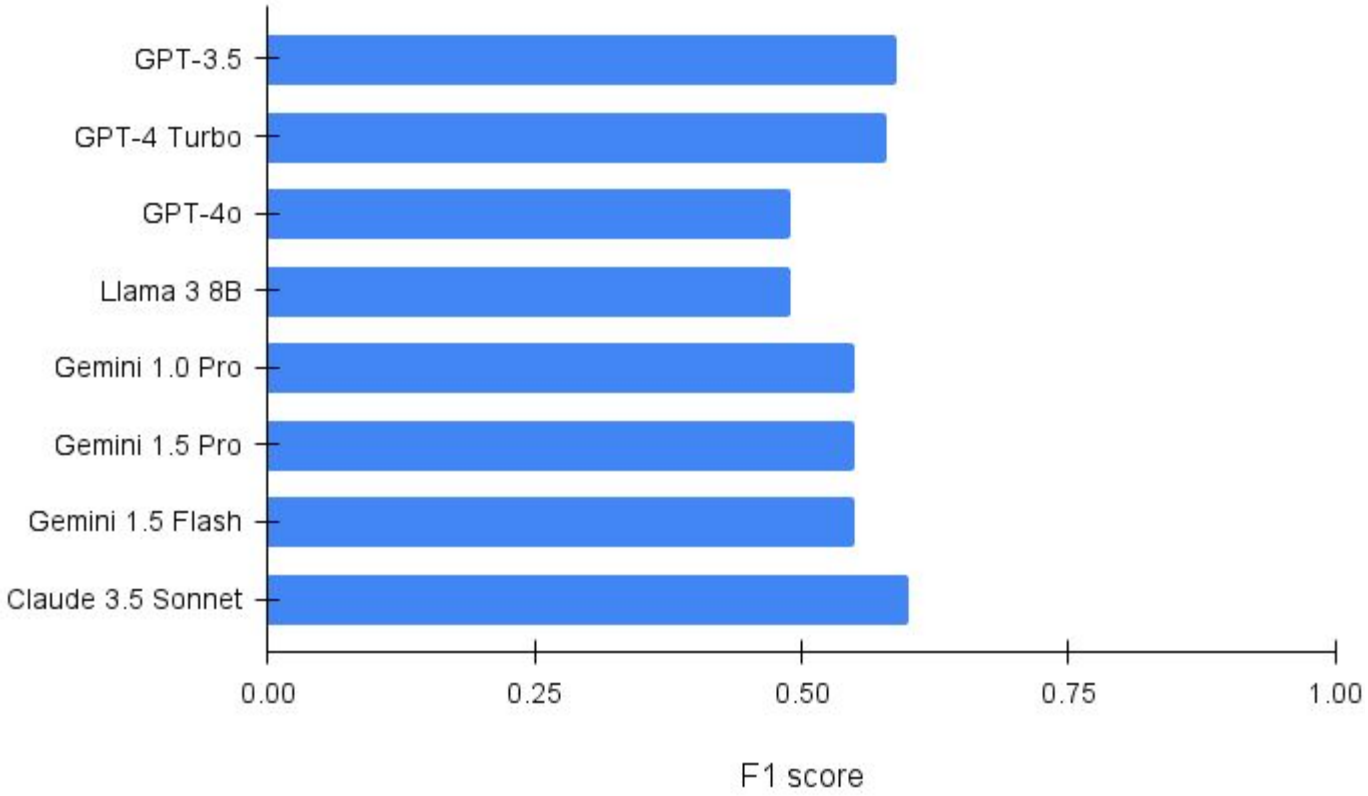


Experiments - Answer Only (A)

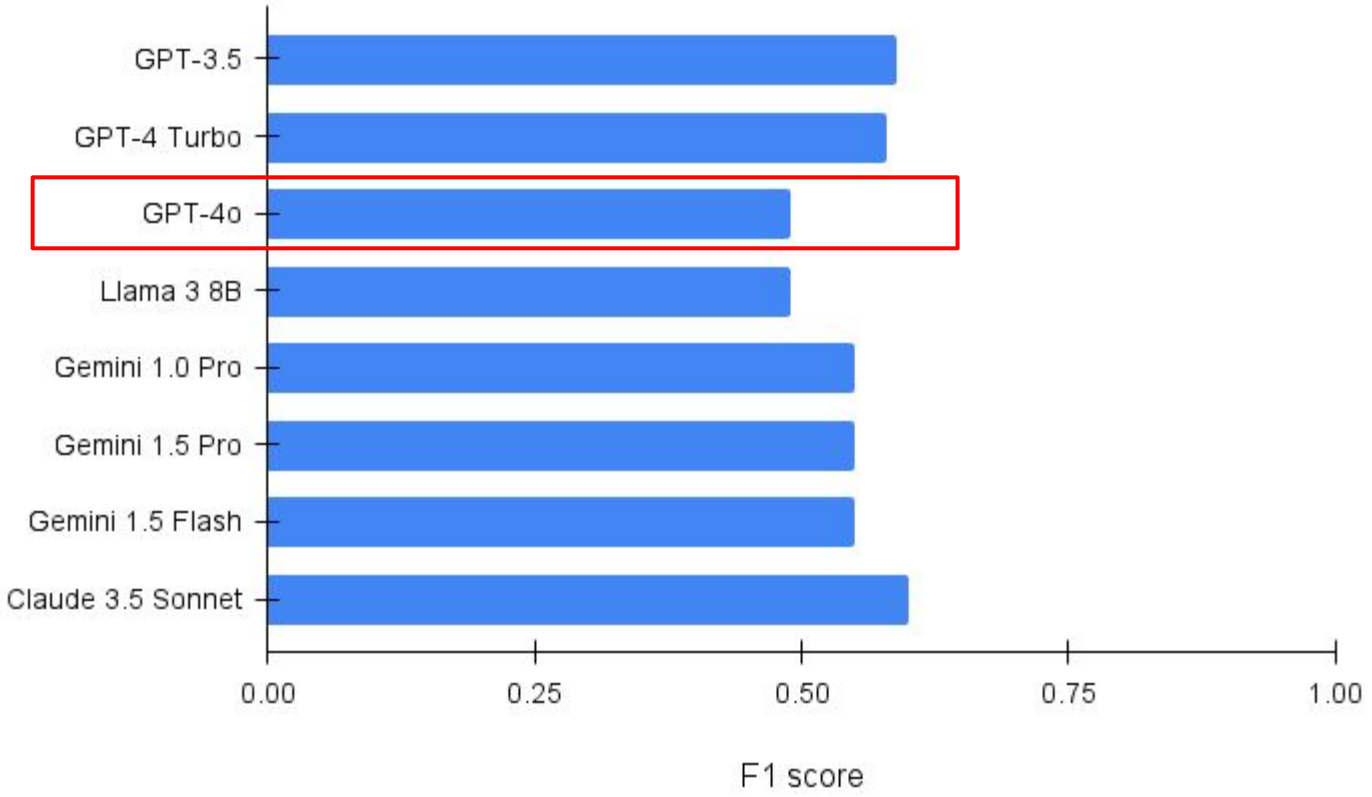
- Binary Prediction

Answer-only (A)	Given a goal, a procedure to achieve that goal and a question about the steps in the procedure, you are required to answer the question in one sentence. Goal: {title} Procedure: {procedure} Must Step {i} happen before Step {j}? Select between yes or no

Models Struggle at Predicting Step Order



Creating 🐱 CaT-Bench

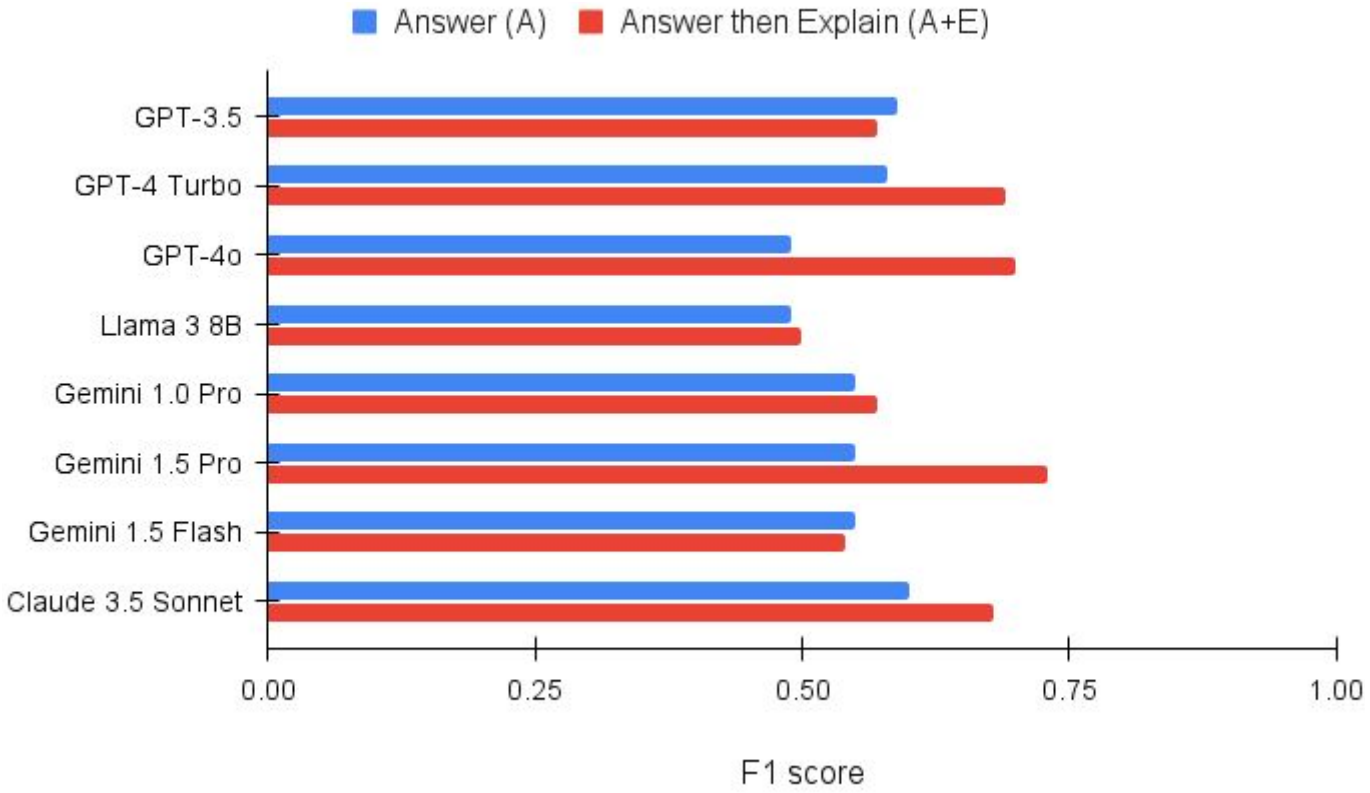


Experiments - Answer + Explanation (A+E)

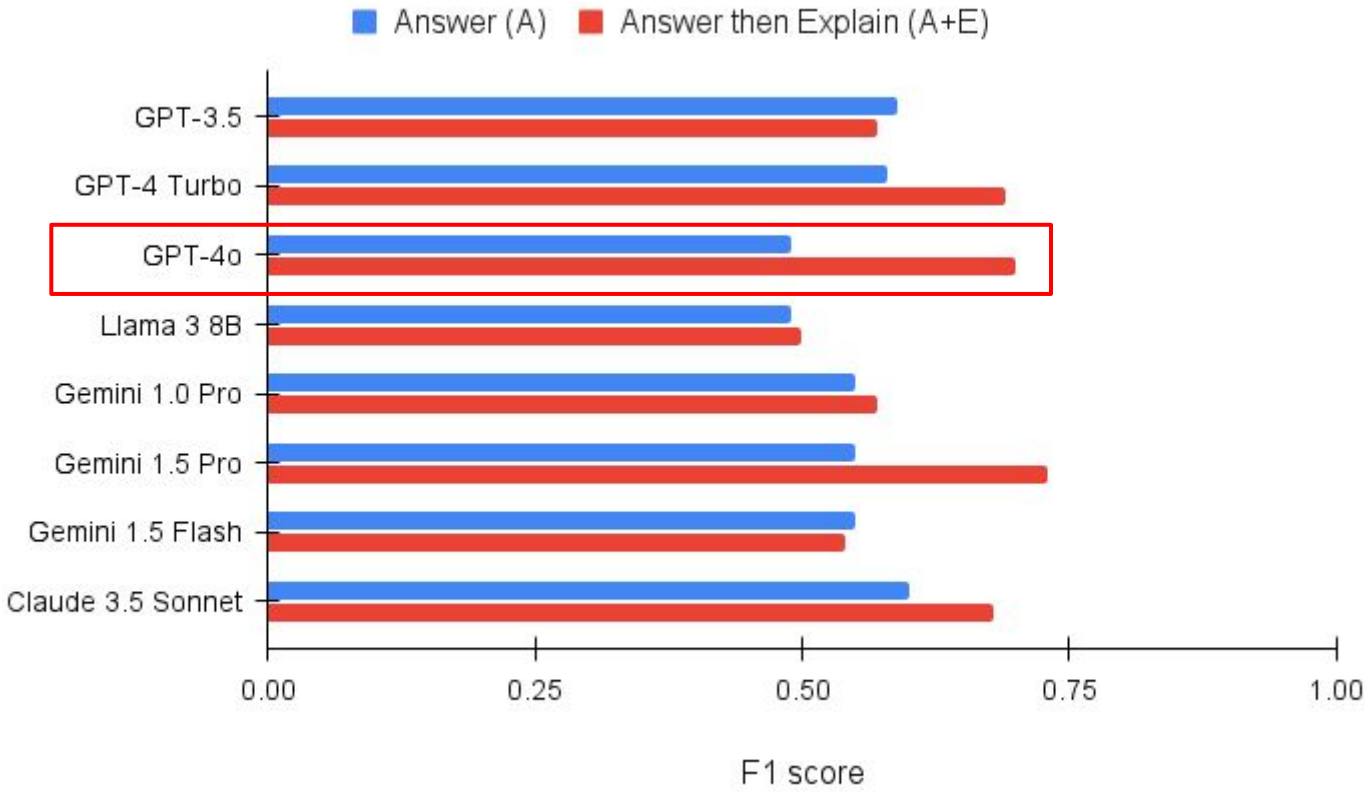
- Binary Prediction and Post-hoc Explanation

Answer + Explanation (A+E)	Given a goal, a procedure to achieve that goal and a question about the steps in the procedure, you are required to answer the question in one sentence. Goal: {title} Procedure: {procedure} 1. Must Step {i} happen before Step {j}? Select between yes or no 2. Explain why or why not. Format your answer as JSON with the key value pairs "binary_answer": "yes/no answer to Q1", "why_answer": "answer to Q2"

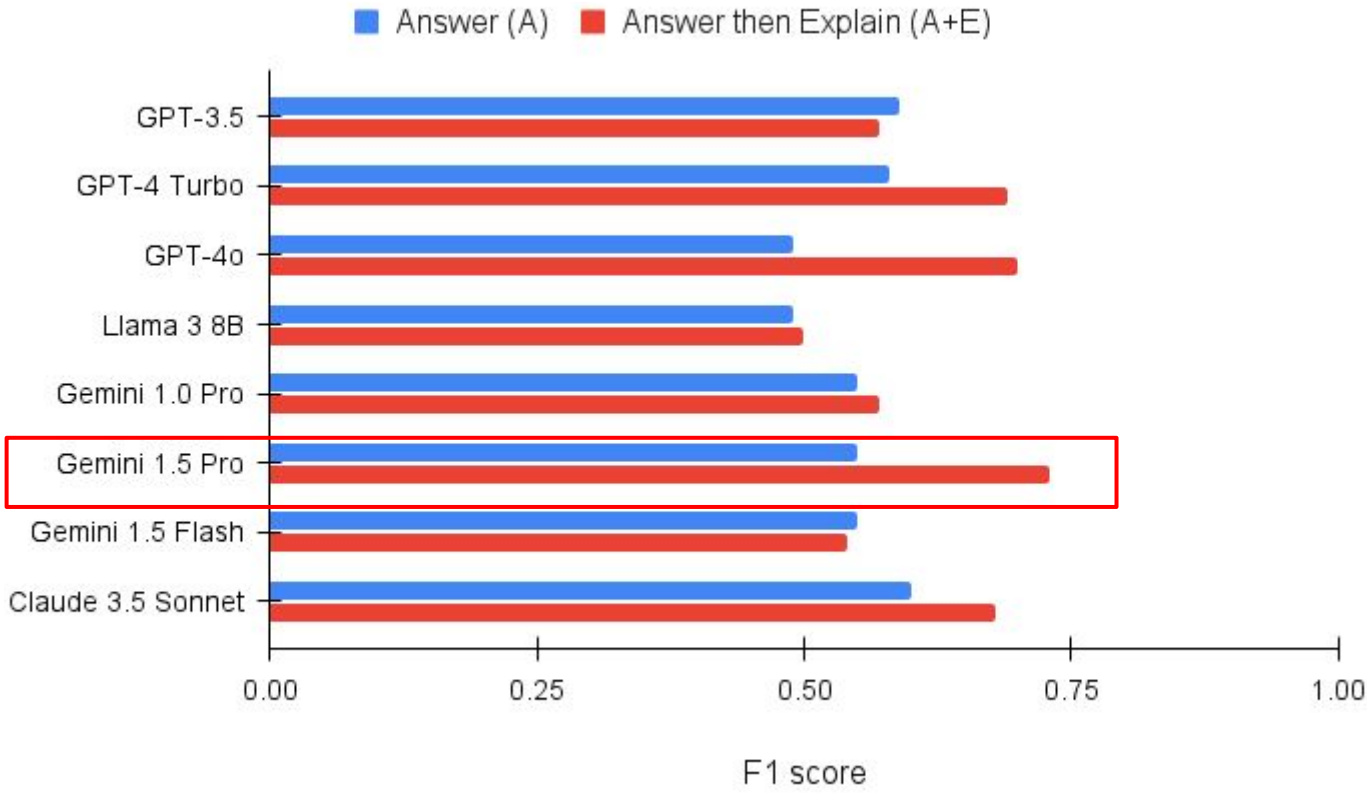
Generating Explanations Helps!



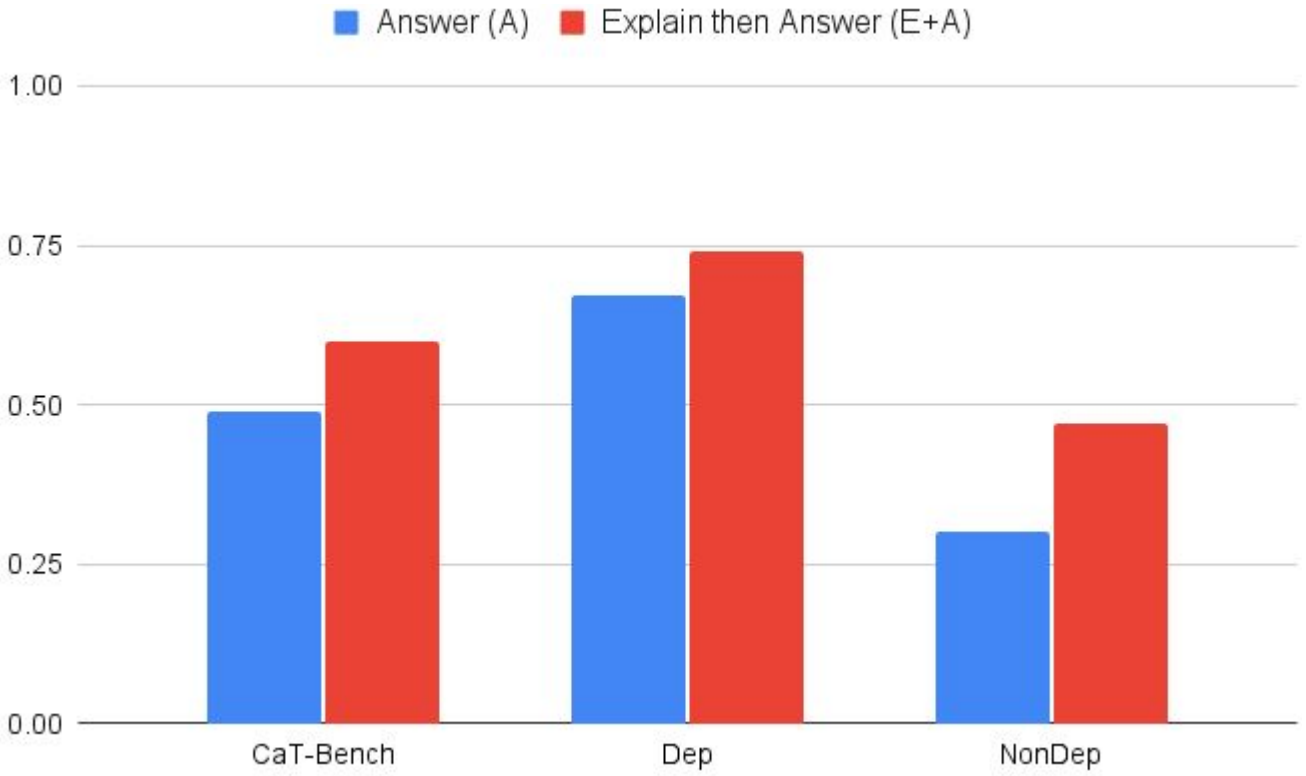
Generating Explanations Helps!



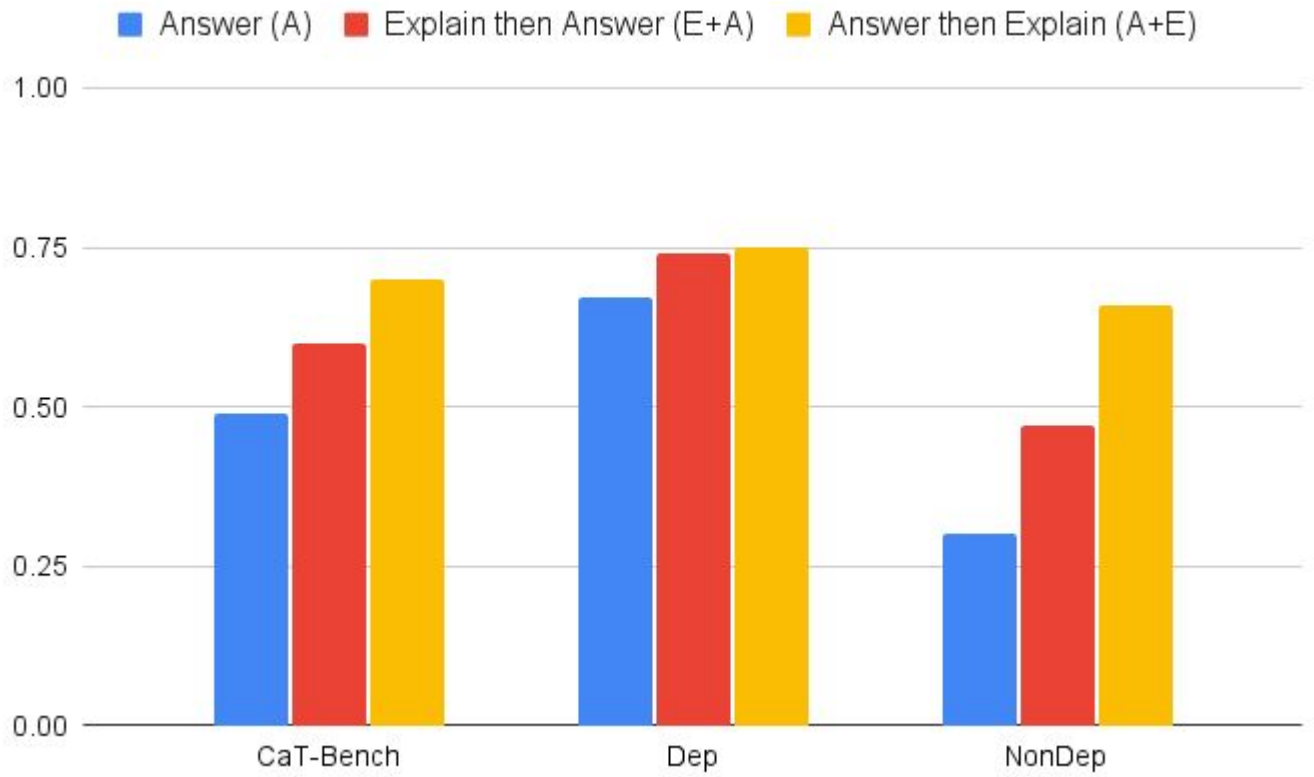
Generating Explanations Helps!



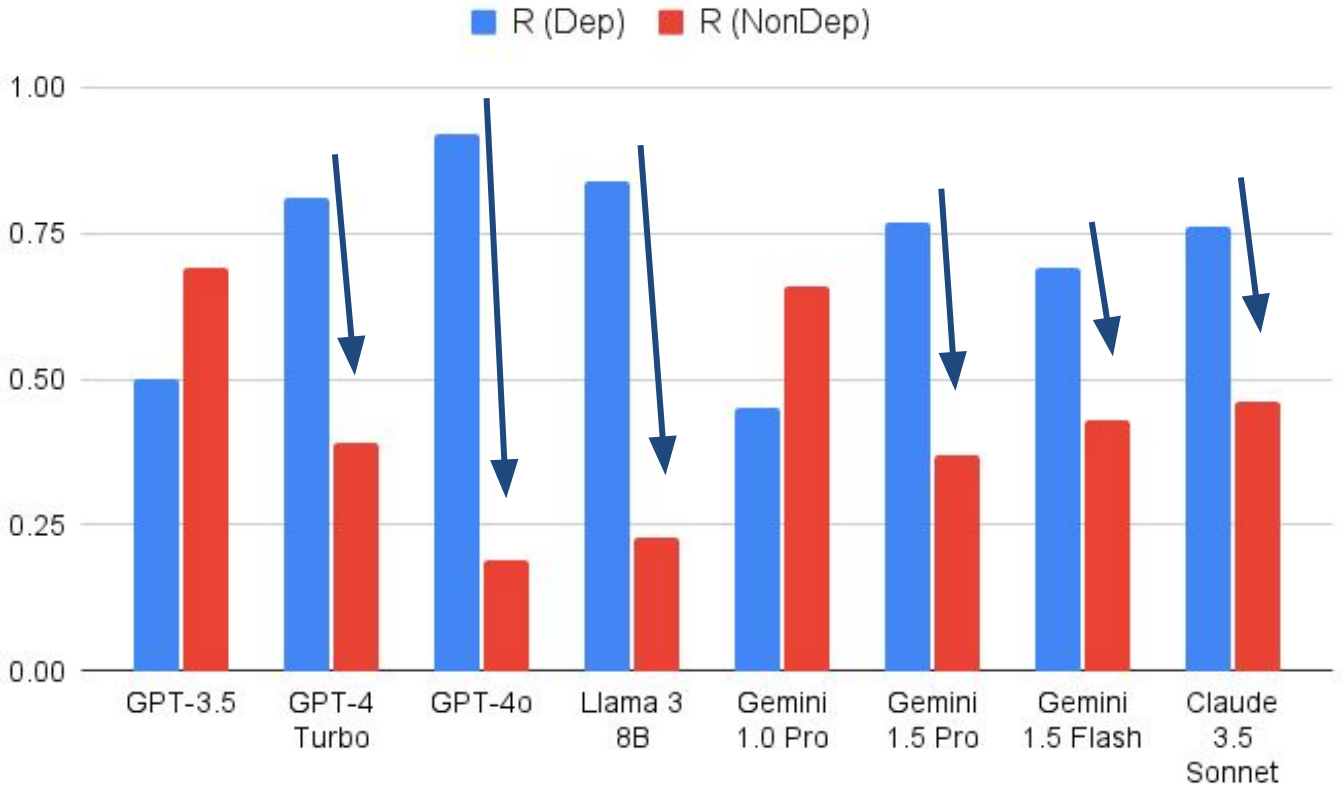
Chain-of-Thought



Post-Hoc Explanation Beats Chain-of-Thought



Models are Biased Towards Dependence



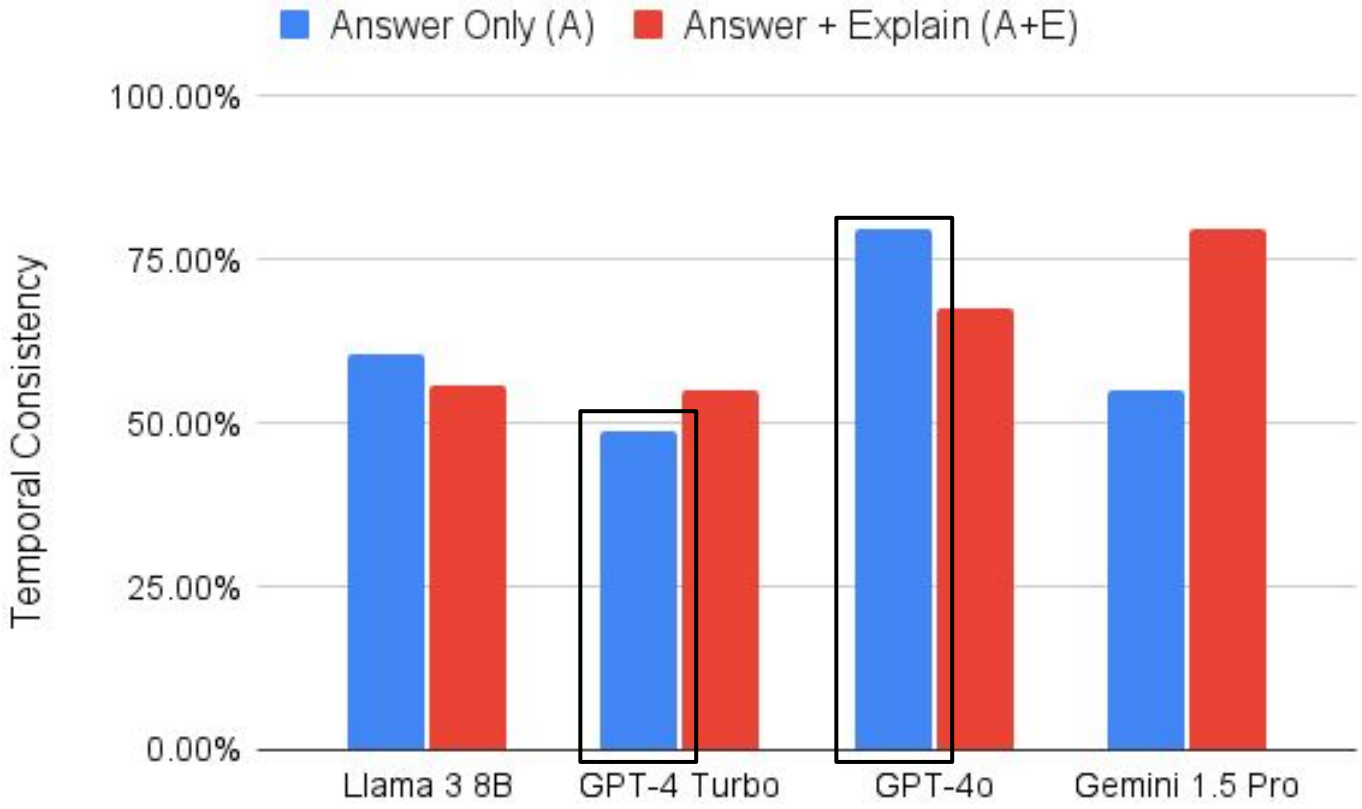
Robustness: Temporal Consistency

Goal: Bake Almond and Chocolate Cake
… Step 6: Add in ground almonds. Step 7: Add half flour and half milk. Step 8: Use wooden spoon to stir. … Step 12: Whip cream till stiff peaks …

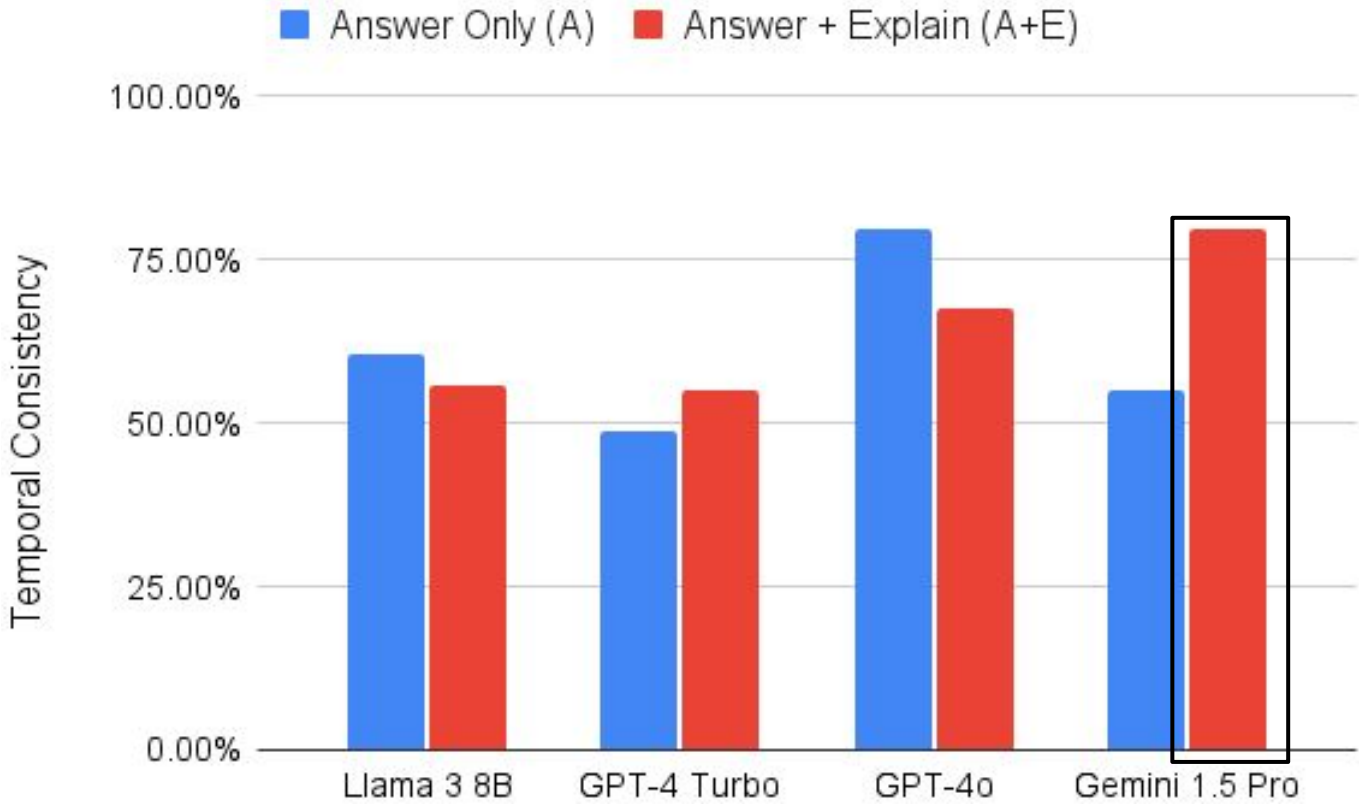
Q: Must Step 6 happen before Step 8?	Before
Q: Must Step 8 happen after Step 6?	After

The model answer to the ‘before’ and ‘after’ questions should be the same.

How Robust are Models?



How Robust are Models?

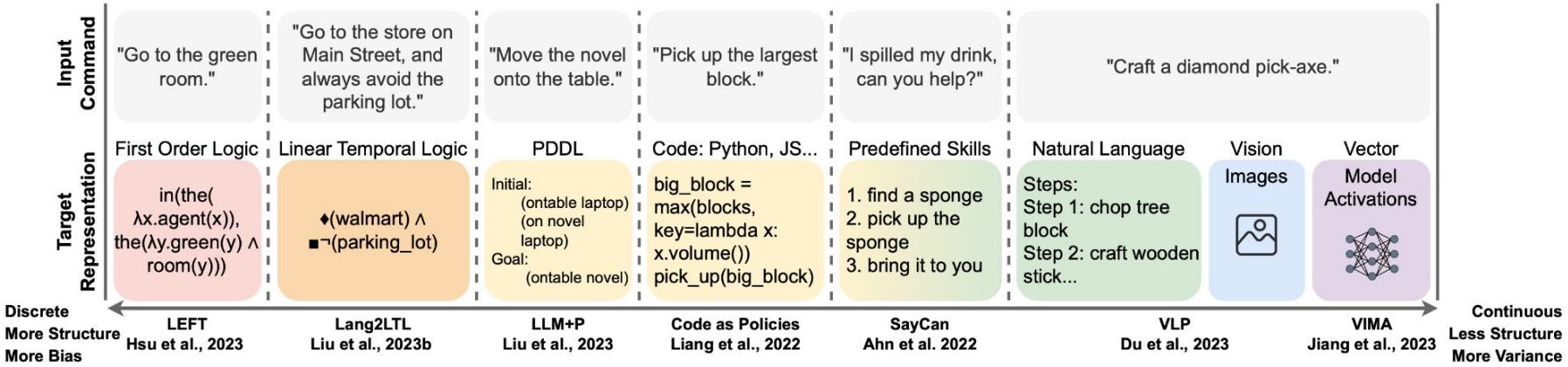


Conclusion

- We introduce an easy-to-evaluate plan based reasoning benchmark.
- SOTA LLMs struggle with this simple task but post-hoc explanations help.
- Models are inconsistent and biased towards predicting step dependence.

A Survey of Robotic Language Grounding: Tradeoffs Between Symbols and Embeddings (IJCAI 2024: Survey Track)

Vanya Cohen*, Jason Xinyu Liu*, Raymond Mooney*, Stefanie Tellex*, David Watkins*



Contents

- Introduction and Related Work
- Completed Work
- **Future Work**
  - Ongoing Work and Short-Term Proposal
  - Long-Term Proposals

MET-Bench: Multimodal Entity Tracking for Evaluating the Limitations of Vision-Language and Reasoning Models

Vanya Cohen and Raymond Mooney

Motivation

- Agents need to track the state of the world in text and visual modalities.
  - Entity state tracking evaluated extensively in text tasks[1].

1. Shubham Toshniwal, Sam Wiseman, Karen Livescu, and Kevin Gimpel. Chess as a Testbed for Language Model State Tracking. In Proceedings of the AAAI 2022
   Najoung Kim and Sebastian Schuste.. Entity Tracking in Language Models. ACL 2023
   Erwan Fagnou, Paul Caillon, Blaise Delattre, and Alexandre Allauzen. Chain and Causal Attention for Efficient Entity Tracking. EMNLP 2024

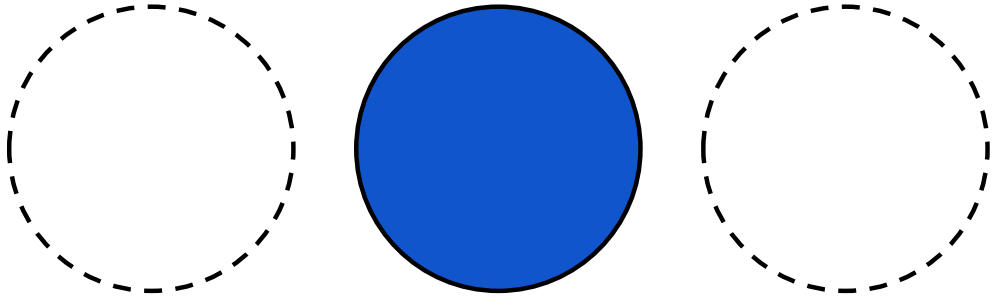
Entity State Tracking

Example: Shell Game

- Ball placed under one of three shells.
- Series of swaps change the hidden ball position.
- Predict the final position of the ball.

Example: Text Shell Game

Ball Position: 2

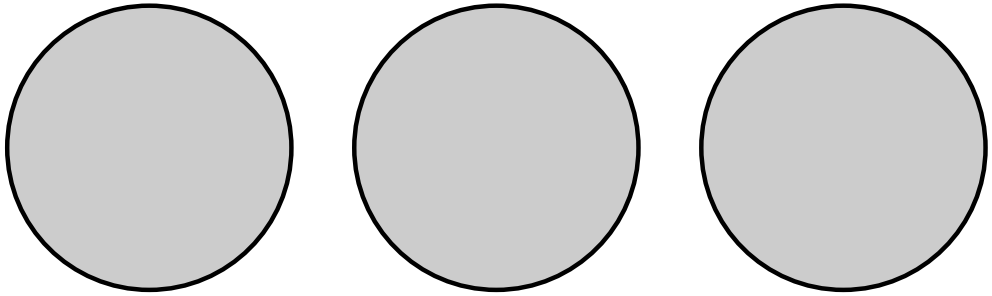


Example: Text Shell Game

Swap 1, 2

Swap 2, 3

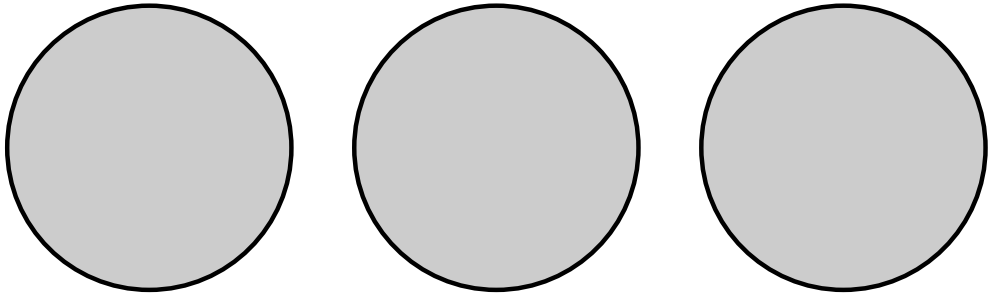
Swap 1, 3



1 2 3

Example: Text Shell Game

Q: Final State?

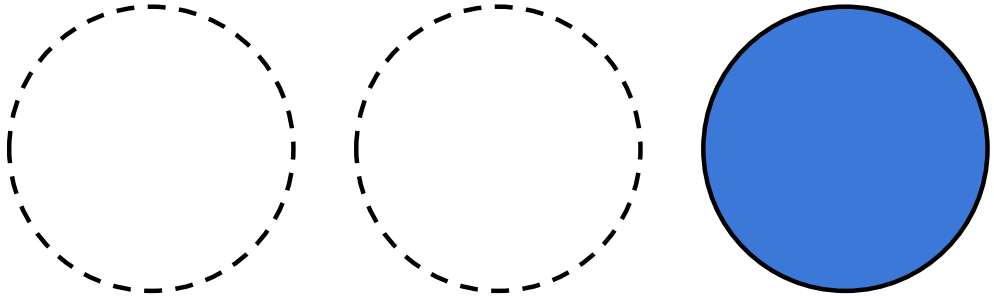


1 2 3

Example: Text Shell Game

Q: Final State?

A: 3



Example: Multimodal Shell Game

Text Input:

Ball Position: 2

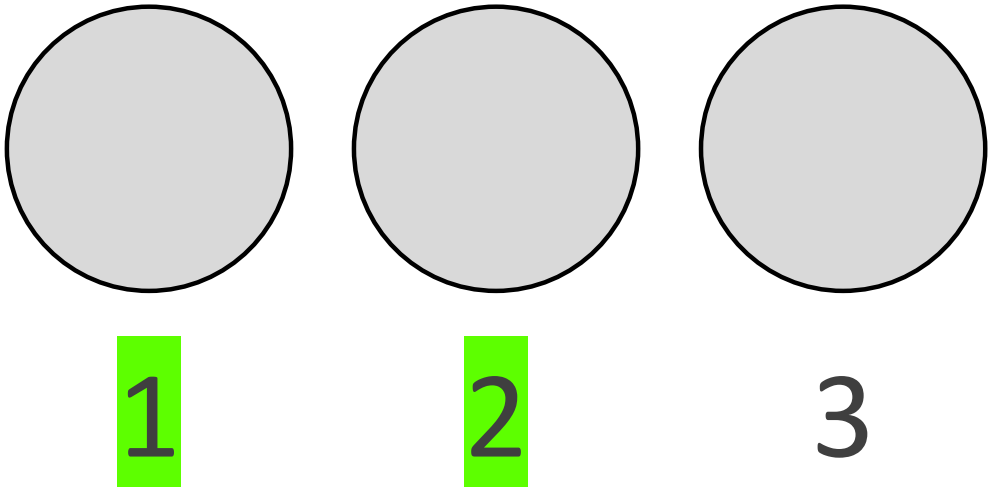
Example: Multimodal Shell Game

Model receives a text description of the initial state

Text Input:

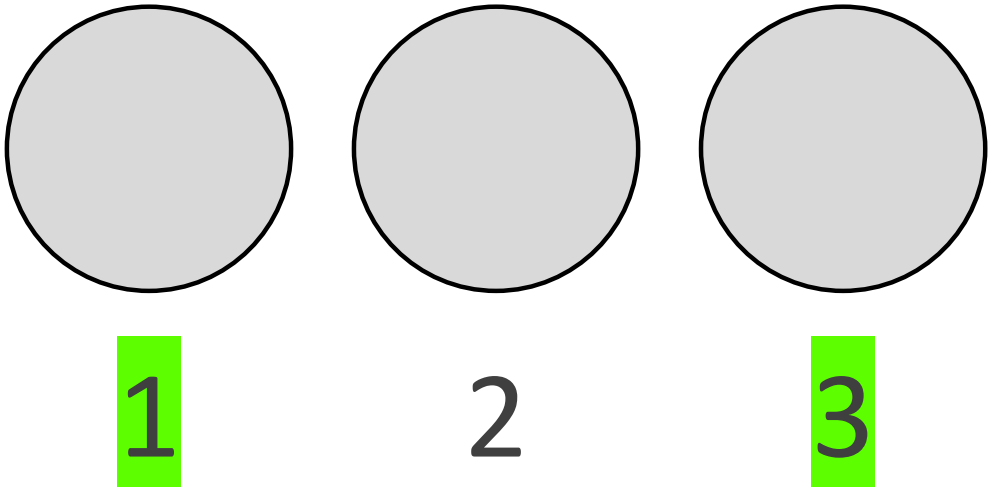
Ball Position: 2

Example: Multimodal Shell Game



Model receives an image depiction of the action

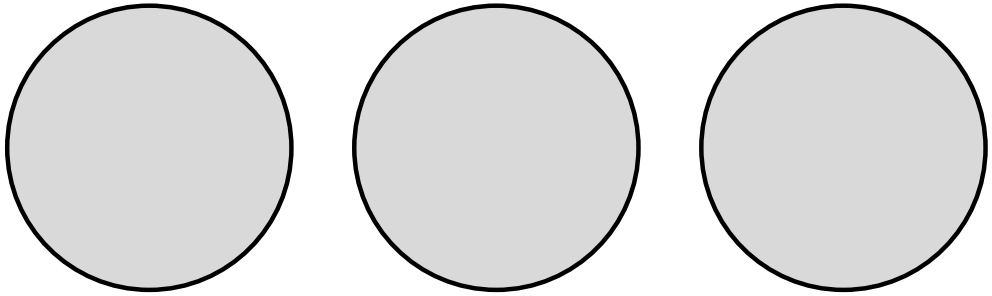
Example: Multimodal Shell Game



Model receives an image depiction of the action

Example: Multimodal Shell Game

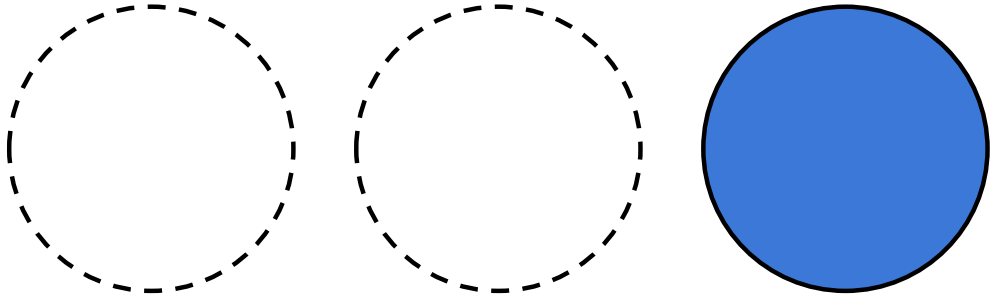
Q: Final State?



Example: Multimodal Shell Game

Q: Final State?

A: 3



Example: Multimodal Chess

Text Input:
Forsyth–Edwards Notation (FEN)

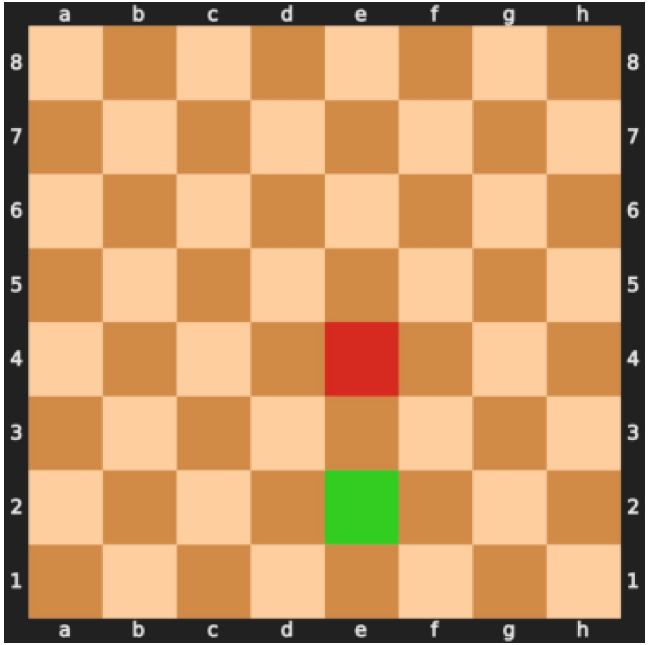
"rnbqkbnr/pppppppp/8/8/8/8/
PPPPPPPP/RNBQKBNR w KQkq -
0 1"

Example: Multimodal Chess

Universal Chess Interface Notation

e2e4

“The piece at e2 moves to e4.”

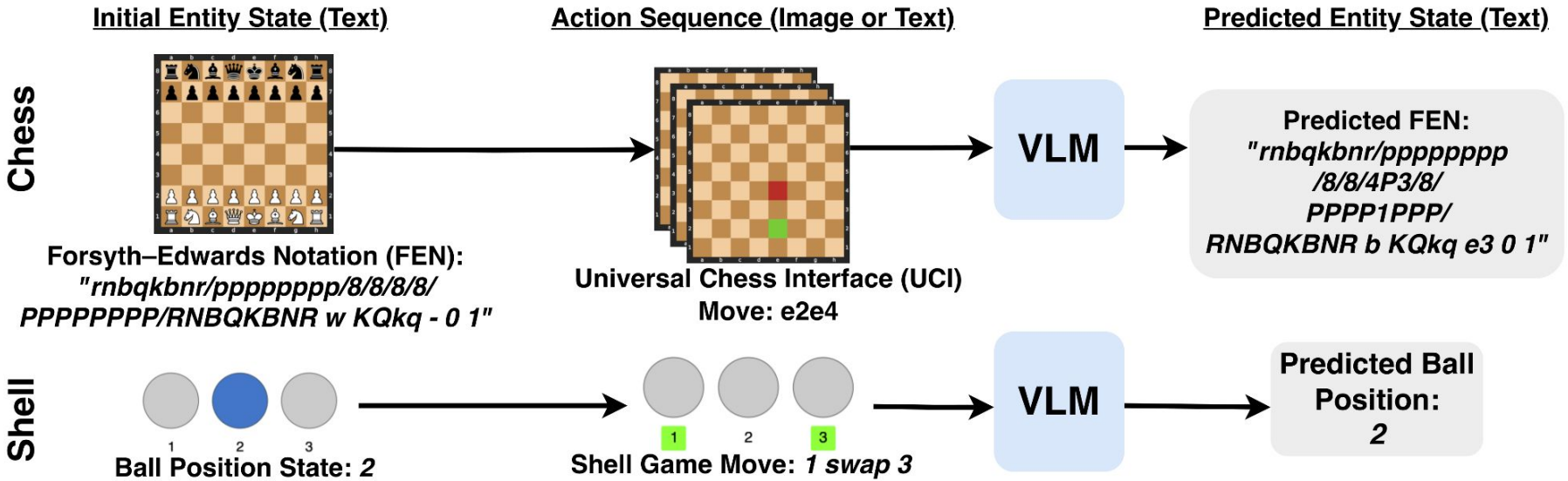


Example: Multimodal Shell Game

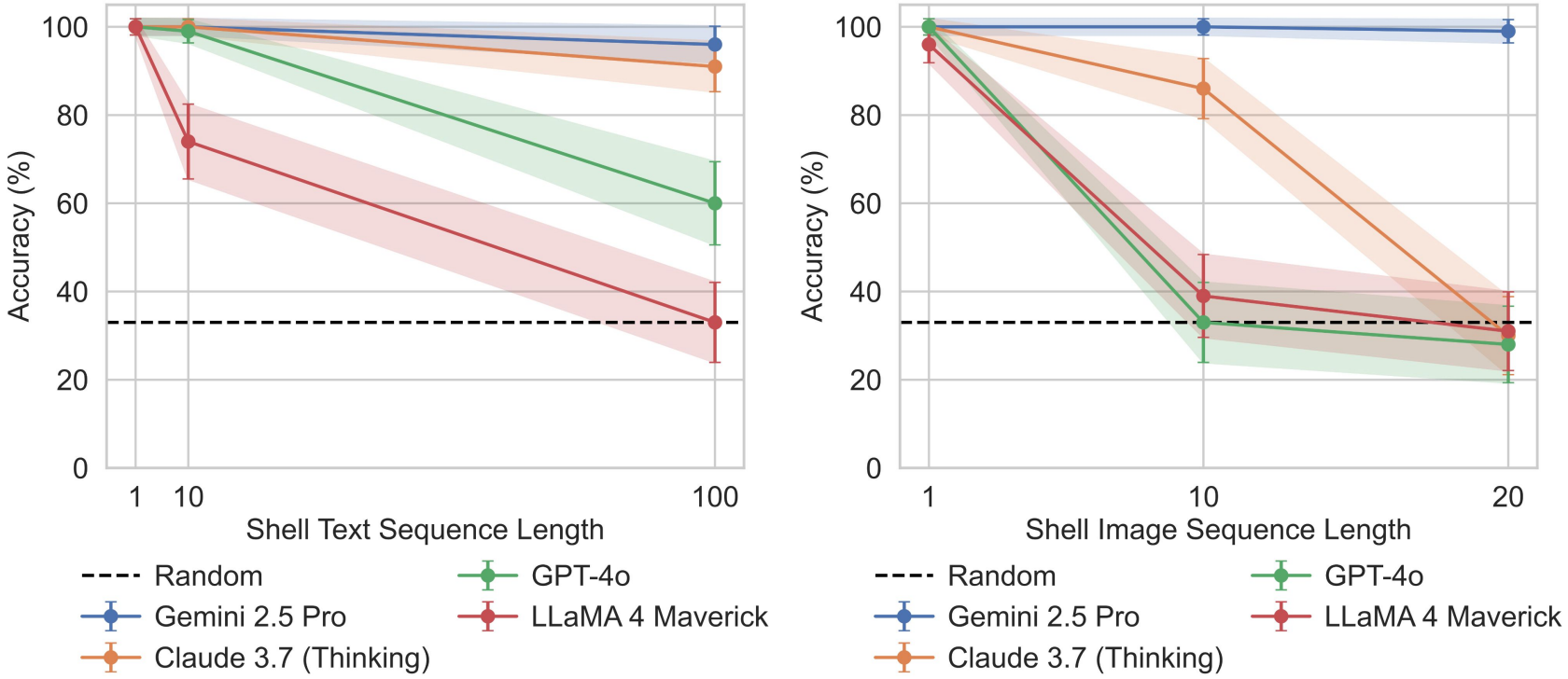
Q: Final State?

A:
rnbqkbnr/pppppppp/8/8/4P3/8/PPPP1PPP/RNBQKBNR b KQkq e3 0 1

MET-Bench: Multimodal Entity State Tracking



Shell Game: Reasoning



Chess: Reasoning

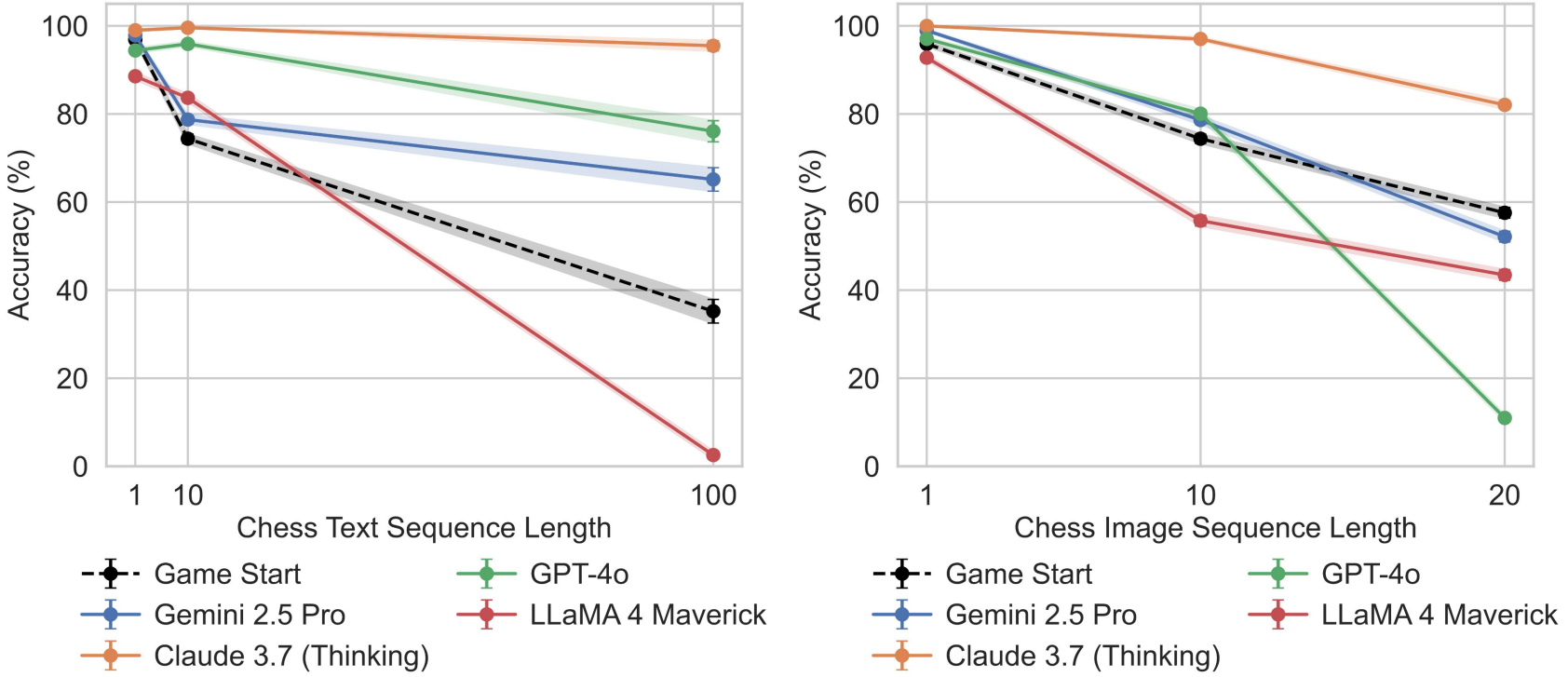
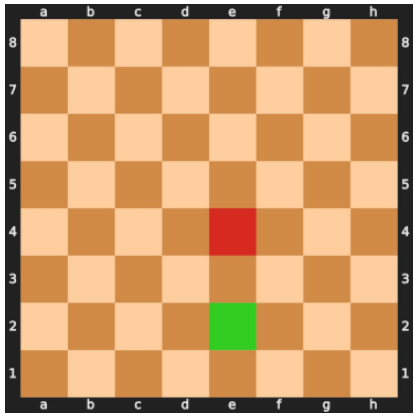


Image Action Perception

Q: In UCI notation what move does the arrow on the chessboard represent? The move is from the green square to the red square. (e.g., 'g2f3').



Q: Which shells are being swapped in the image? Shells are labeled '1', '2', '3' and the shells being swapped have their numbers highlighted in green. Only output a move in the format 'X swap Y' and nothing else.

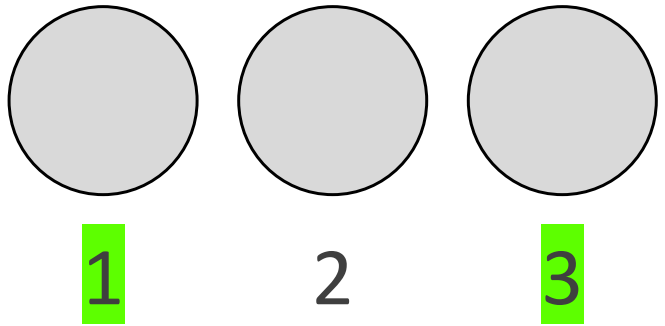


Image Action Perception

Model	Overall (%)
Chess	
GPT-4o-mini	46.4 ± 1.0
GPT-4o	95.2 ± 0.4
Shell Game	
GPT-4o-mini	100.0 - 0.2
GPT-4o	100.0 - 0.2

Image Action Perception

Model	Overall (%)
Chess	
GPT-4o-mini	46.4 ± 1.0
GPT-4o	95.2 ± 0.4
Shell Game	
GPT-4o-mini	100.0 - 0.2
GPT-4o	100.0 - 0.2

Cascaded Inference

Image Actions → Text Actions → State Prediction

METHOD	CHESS	SHELL
CHAIN-OF-THOUGHT		
GPT-4O MINI	66.4 ± 0.4	47.6 ± 3.1
GPT-4O	93.2 ± 0.2	99.4 ± 0.5

Cascaded Inference

Image Actions → Text Actions → State Prediction

METHOD	CHESS	SHELL
CHAIN-OF-THOUGHT		
GPT-4O MINI	66.4 ± 0.4	47.6 ± 3.1
GPT-4O	93.2 ± 0.2	99.4 ± 0.5

Discussion

- Models perceive the image depictions of actions, but cannot track entity state.
- Reasoning models perform better, especially for long-context and tracking from images.

Limitations

- Synthetic tasks; strategy games.
- Current challenges are in real-world and embodied domains.

Short-Term Proposal

- Expand MET-Bench to embodied tasks.

Short-Term Proposal

- Expand MET-Bench to embodied tasks.
- Human annotation of entity tracking reasoning traces.
  - Evaluate reasoning and collect correct reasoning examples

Short-Term Proposal

Minecraft:

- Challenging benchmark for planning and acting with frontier models[1].
- Semantically rich, varied tasks.
- Accessible world state.

1. Voyager (Wang et al. 2024); Dreamer 4 (Hafner et al. 2025)

Perception

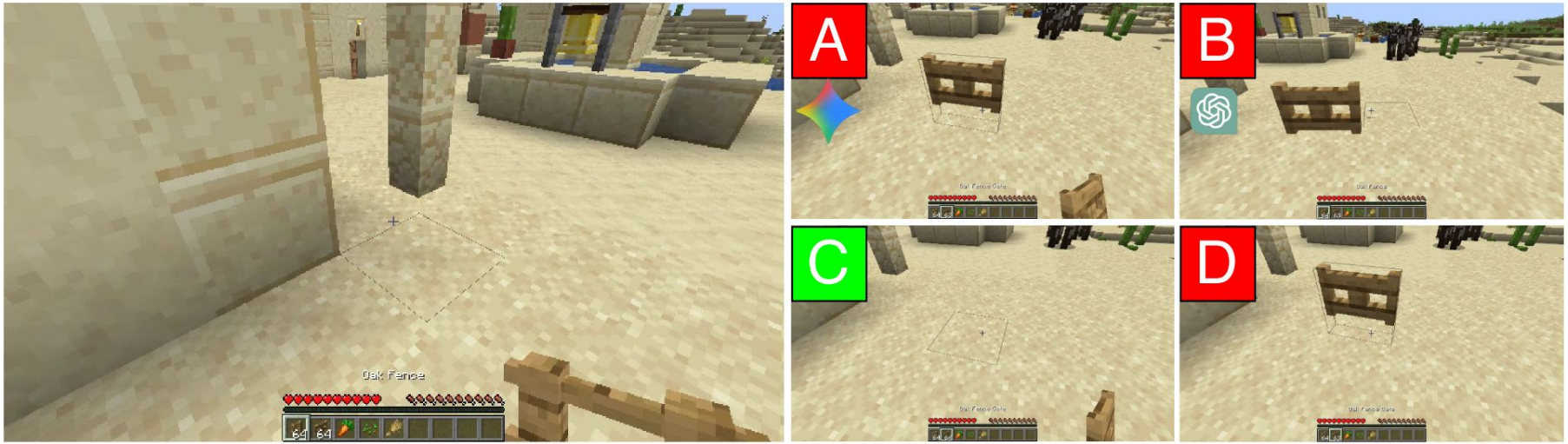
Q: How many villagers are in the image?

Perception

Q: How many villagers are in the image?

World Modeling (State Tracking)

Q: The player turns right. Which is the most likely view?



Reasoning

Reasoning

Q: The player turns right. Which is the most likely view?

Reasoning

Reasoning

Q: The player turns right. Which is the most likely view?

Comparing the images, the most likely result of rotating right from your starting image is Candidate B. Here's why:

- The pillar and building wall from the starting view are now on the left, which makes sense after a right turn.
- The cows and cacti come into view directly ahead — they weren't visible before, but they are in this direction relative to the initial position.
- The lighting and angle align perfectly with a small rightward rotation from the starting point.

Reasoning

Reasoning

Q: The player turns right. Which is the most likely view?

Comparing the images, the most likely result of rotating right from your starting image is Candidate B. Here's why:

- The pillar and building wall from the starting view are now on the left, which makes sense after a right turn.
- The cows and cacti come into view directly ahead — they weren't visible before, but they are in this direction relative to the initial position.
- The lighting and angle align perfectly with a small rightward rotation from the starting point.

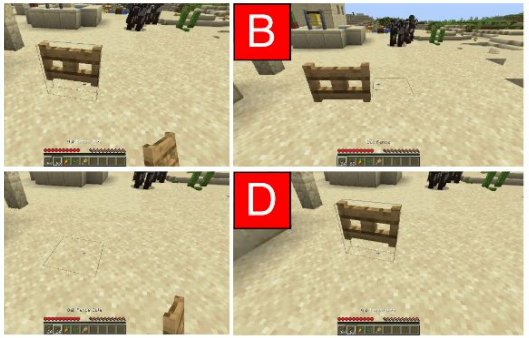
Reasoning

Reasoning

Q: The player turns right. Which is the most likely vi

Comparing the images, the most likely result of rotating right from your starting image is Candidate B. Here's why:

- The pillar and building wall from the starting view are now on the left, which makes sense after a right turn.
- The cows and cacti come into view directly ahead — they weren't visible before, but they are in this direction relative to the initial position.
- The lighting and angle align perfectly with a small rightward rotation from the starting point.



Reasoning

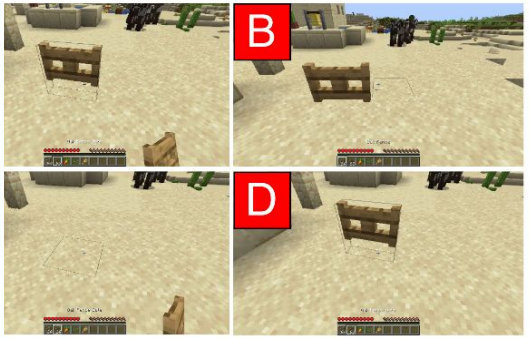
Reasoning

Q: The player turns right. Which is the most likely vi

Comparing the images, the most likely result of rotating right from your starting image is Candidate B. Here's why:

X

- The pillar and building wall from the starting view are now on the left, which makes sense after a right turn.
- The cows and cacti come into view directly ahead — they weren't visible before, but they are in this direction relative to the initial position.
- The lighting and angle align perfectly with a small rightward rotation from the starting point.



Reasoning

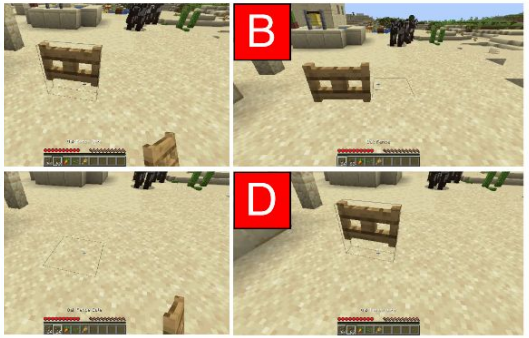
Reasoning

Q: The player turns right. Which is the most likely vi

Comparing the images, the most likely result of rotating right from your starting image is Candidate B. Here's why:

X
- The pillar and building wall from the starting view are now on the left, which makes sense after a right turn.

X
- The cows and cacti come into view directly ahead — they weren't visible before, but they are in this direction relative to the initial position.
- The lighting and angle align perfectly with a small rightward rotation from the starting point.



Reasoning

Reasoning

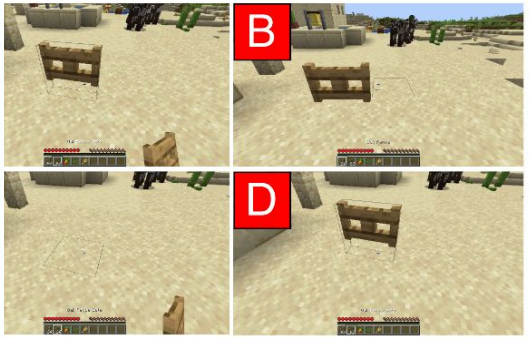
Q: The player turns right. Which is the most likely vi

Comparing the images, the most likely result of rotating right from your starting image is Candidate B. Here's why:

X • The pillar and building wall from the starting view are now on the left, which makes sense after a right turn.

X • The cows and cacti come into view directly ahead — they weren't visible before, but they are in this direction relative to the initial position.

X • The lighting and angle align perfectly with a small rightward rotation from the starting point.



Long-Term Proposals

- CaT-Bench and MET-Bench
  - Reasoning models perform best at these tasks.
- CoT traces show brittle reasoning.
  - CaT-Bench CoT underperforms.

Long-Term Proposals

- RL to improve improve MET-Bench and CaT-Bench-type task reasoning.
  - Utilize human-annotated reasoning traces to bootstrap learning.

Long-Term Proposals

Reasoning

Q: The player turns right. Which is the most likely view?

Comparing the images, the most likely result of rotating right from your starting image is Candidate B. Here's why:

- The pillar and building wall from the starting view are now on the left, which makes sense after a right turn.
- The cows and cacti come into view directly ahead — they weren't visible before, but they are in this direction relative to the initial position.
- The lighting and angle align perfectly with a small rightward rotation from the starting point.

Reward Signal

Long-Term Proposals

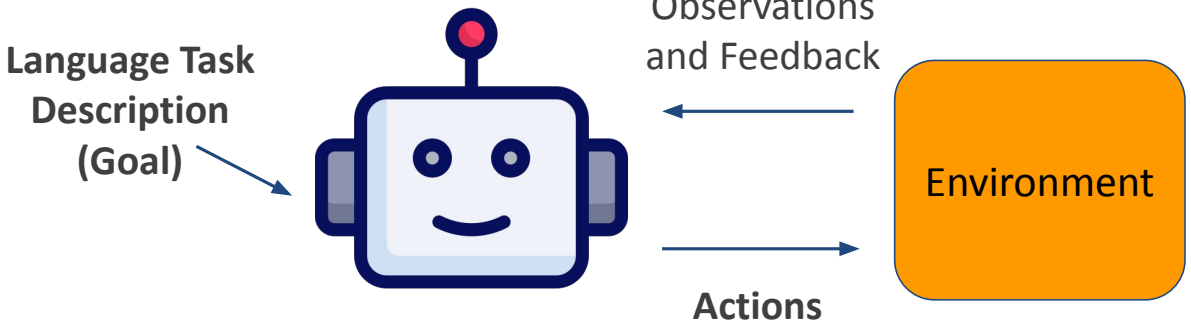
- Test action reasoning as an auxiliary RL task for improving agentic instruction following.

Auxiliary Action-Reasoning Training with RL

Preconditions
CaT-Bench

Postconditions
CaT-Bench
MET-Bench

Goal-Action
Home robotics, embodied simulation.



Produce **valid actions** to **accomplish a task** specified in **natural language**.

Proposed Timeline

Nov - Dec:

- Collect embodied task dataset in Minecraft.

Jan - Feb:

- Human evaluation of MET-Bench model reasoning traces.
- Collect human ground-truth reasoning traces.

March:

- Initial results:
  - Evaluate finetuning on reasoning traces.
  - RL training for improving reasoning for MET-Bench tasks.

Defense in mid April.

Questions and Discussion