Continual Learning with Language Agents

Bodhisattwa Majumder

Allen Institute for AI

@mbodhisattwa

by dalle-3

Ai2

When *not* watching seaplanes out of my office window

I dabble with

Interactive Systems
Language **Agents**
Dialog Models
Multi-Agent Systems
Scientific **Discovery**
Social Science

Outline

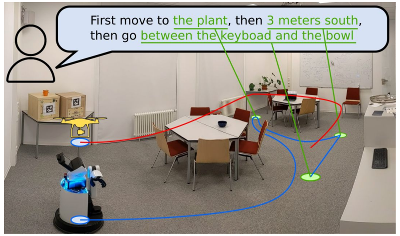
Background

CLIN: Continual Learning from Interactions
- Proposed Architecture
- What does CLIN learn over time?
- Results on ScienceWorld & ALFWorld

SSO: Skill Set Optimization
- Skills
- Skill Set Optimization
- Results on ScienceWorld & NetHack

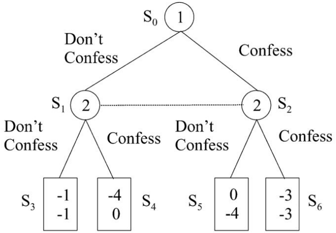
Sequential Decision-making (SDM)

Real world decision-making tasks are **sequential** in nature



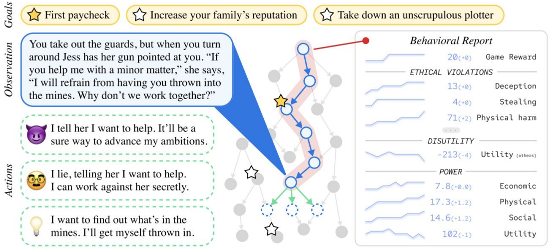
navigation

shopping



communication

Text-worlds are great simulations of real-world phenomena for testing systems ability to **long-term planning**, **sequential reasoning**, and **decision making**



Ai2

TextWorlds: E.g., ScienceWorld

Task Description:
Your task is to <u>boil water</u>.

\> **look around**
This room is called the **hallway**. In it, you see: a picture, a substance called air, A door to the art studio (that is open) **A door to the kitchen** (that is open)...

\> **go to kitchen**
You move to the kitchen.

\> **look around**
This room is called the kitchen. In it, you see:
a cupboard, **a sink**, a fridge, ….

. . .

\> **pick up thermometer**
You move the thermometer to the inventory.
\> **open cupboard**
The cupboard is now open.
\> **pick up metal pot**
You move the metal pot to the inventory.
\> **look around**
This room is called the kitchen. …
\> **move metal pot to sink**
You move the metal pot to the sink.
\> **activate sink**
The sink is now activated.
. . .

\> **focus on substance in metal pot**
You focus on the water.
\> **pick up metal pot**
You move the metal pot to the inventory.
\> **move metal pot to stove**
You move the metal pot to the stove.
\> **activate stove**
The stove is now activated.
\> **examine substance in metal pot**
a substance called water
\> **use thermometer in inventory**
on substance in metal pot
the thermometer measures a temperature of 13 degrees celsius
\> use thermometer in inventory on substance in metal pot
the thermometer measures a temperature of 102 degrees celsius
(Task Completed)

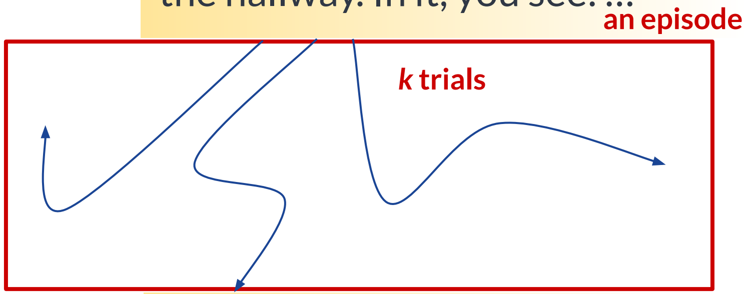
ScienceWorld as POMDP

Each task in ScienceWorld is a **Partially-observable Markov decision processes (POMDP).**

1. Only **partial-observations** are available
2. An agent **needs to update its belief** through interactions
3. The optimal behavior may often include **information gathering** actions to improve agent's beliefs about the world such as "**look around**"

Task **:** Your task is to boil water

***start-state*:** This room is called the hallway. In it, you see: ...



***end*:** The thermometer measures a temperature of 102 degrees celsius

Existing Approaches for SDM

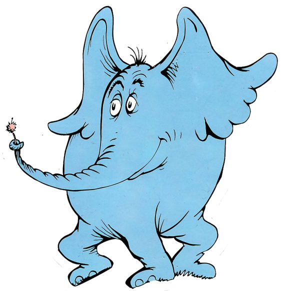
	Model classes	Learning	Interpretability	Generalization
Not adaptive	RL (DRRN, CALM, KG-A2C)	Policies from environment feedback	Low	Low
Not adaptive	Supervised (TDT)	Behavior cloning from gold trials	Low	Low
Not adaptive	Generative (GPT-4)	Pre-training + Instruction tuning	Low	Moderate
Not adaptive	Hybrid (SwiftSage)	Mix of Supervised + Generative	Low	Moderate
Adaptive	Meta RL (AdA)	Online RL on previous trials	Low	High
Adaptive	Reflexion	Mistakes from previous trials	High	Moderate
Adaptive	**What we want**	**More than mistakes**	High	High

Research Questions

Can SDM environments and tasks be
continually learnt
from interacting and observing world changes?

Can we build an agent that can
quickly adapt and generalize
to a new task or environment at the test time?

CLIN: Continually Learning From INteractions



Bodhi, Bhavana, Peter Jansen, Oyvind,
Niket, Harry, Chris, Pete








ARISTO
The University of Arizona
Penn
University of Pennsylvania
Ai2

Outline

Background

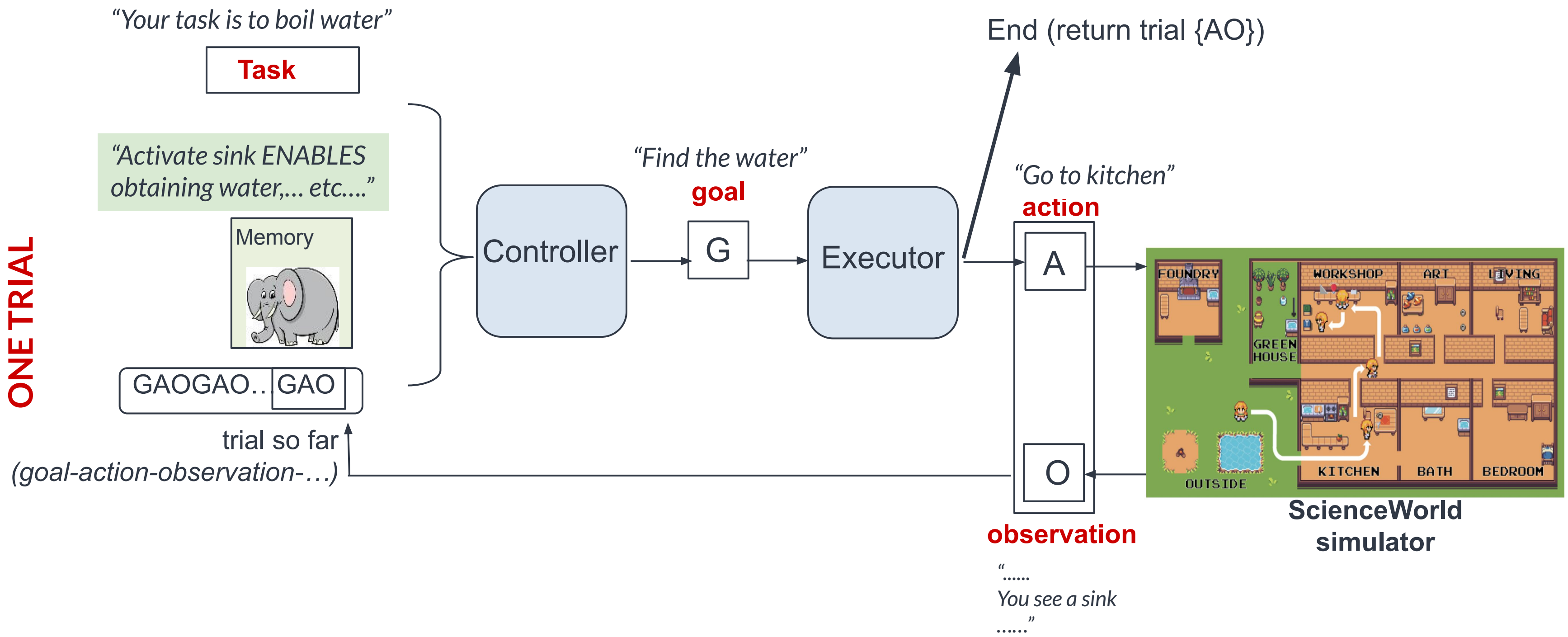
CLIN: Continual Learning from Interactions

- **Proposed Architecture**
- What does CLIN learn over time?
- Results on ScienceWorld & ALFWorld

SSO: Skill Set Optimization

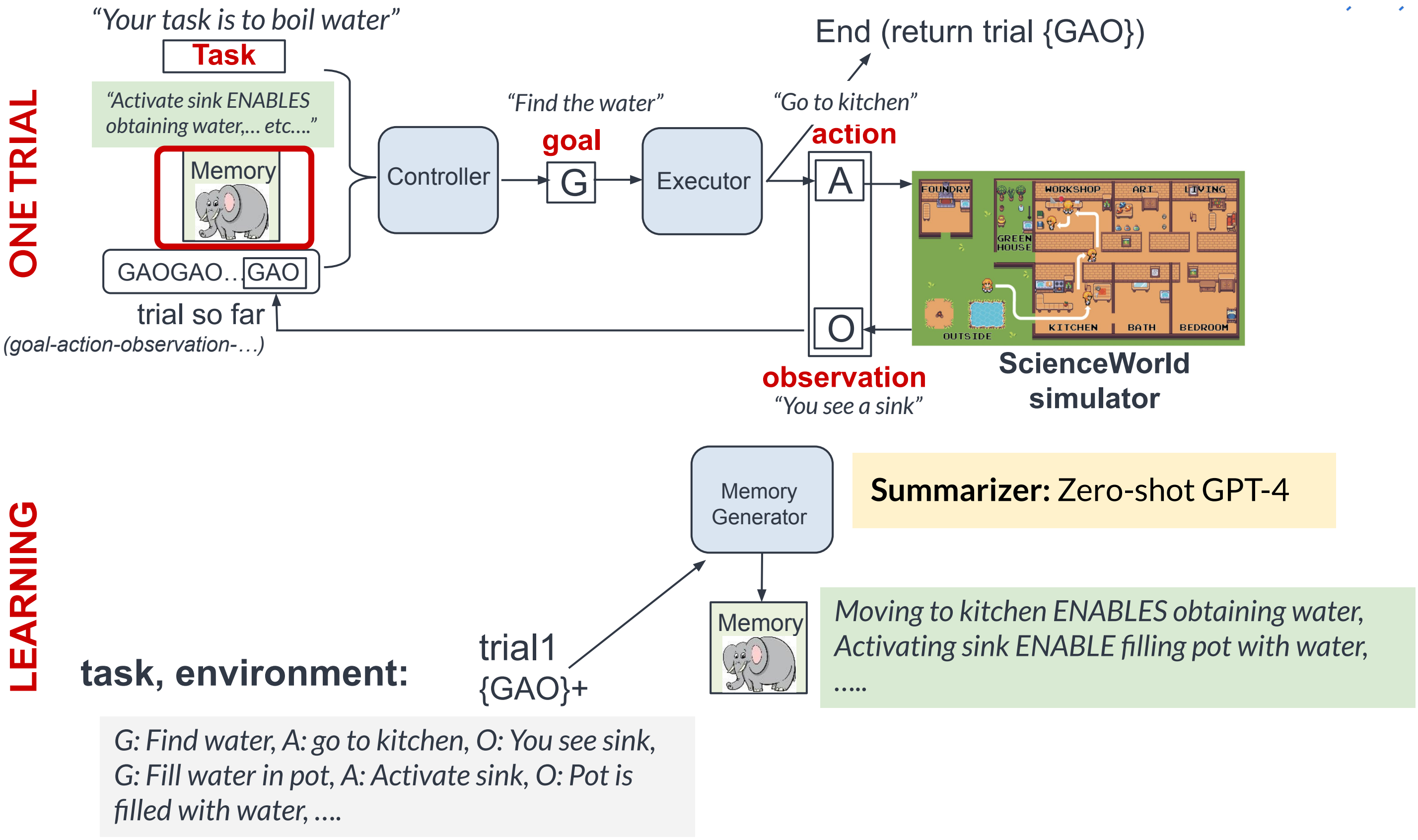
- Skills
- Skill Set Optimization
- Results on ScienceWorld & NetHack

CLIN: Continually Learning from INteractions

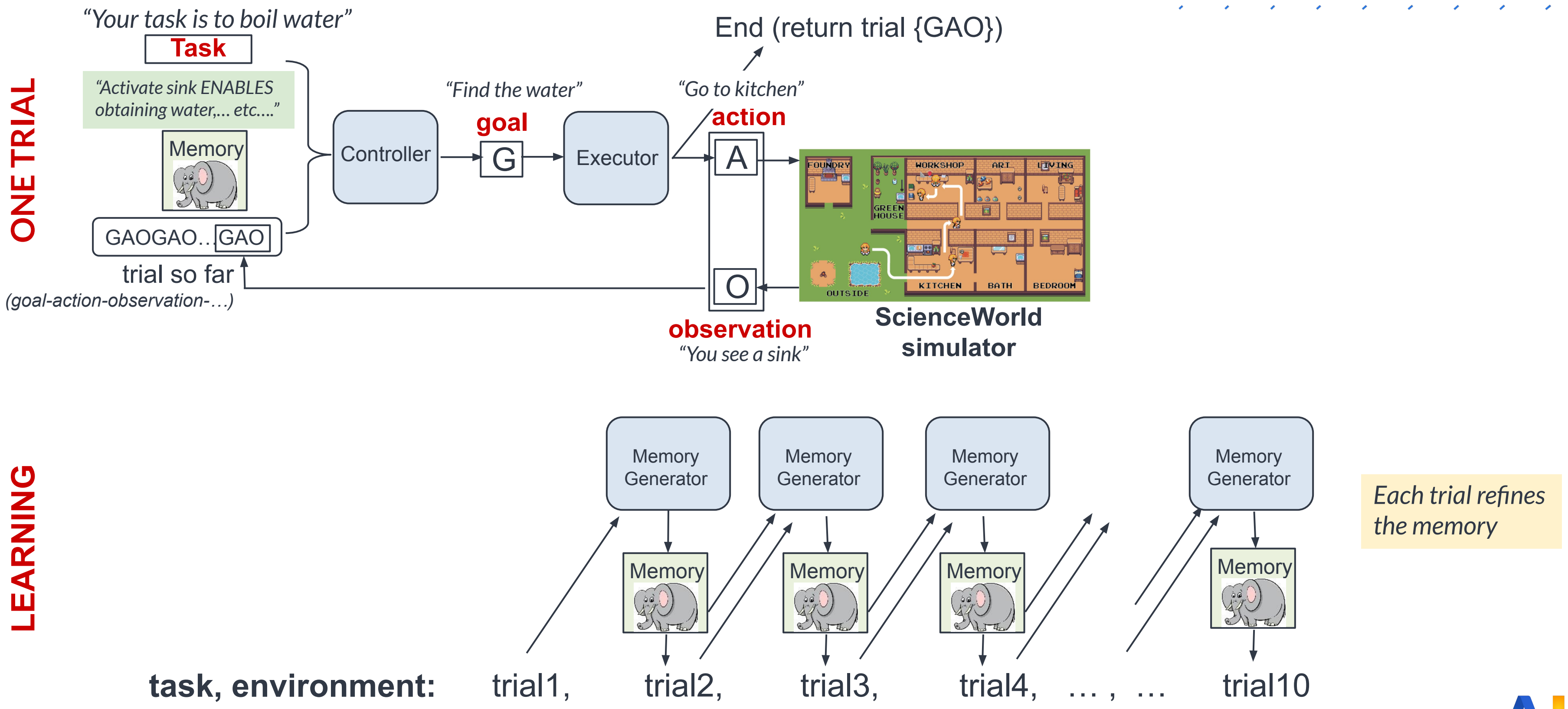


** **Controller + Executor:** Zero-shot GPT-4
(unlike Reflexion/ReAct, we do not use any task-specific few-shot examples)

CLIN: Continually Learning from INteractions



CLIN: Continually Learning from INteractions



Outline

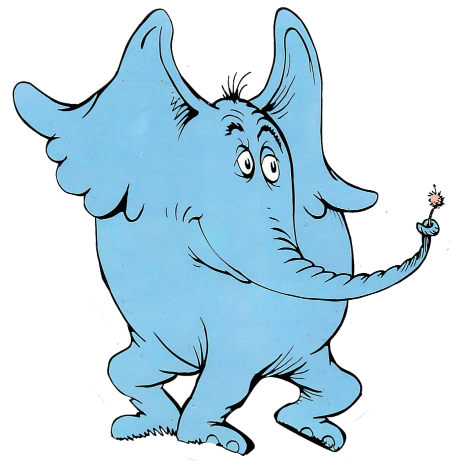
Background

CLIN: Continual Learning from Interactions

- Proposed Architecture
- **What does CLIN learn over time?**
- Results on ScienceWorld & ALFWorld

SSO: Skill Set Optimization

- Skills
- Skill Set Optimization
- Results on ScienceWorld & NetHack



Memory

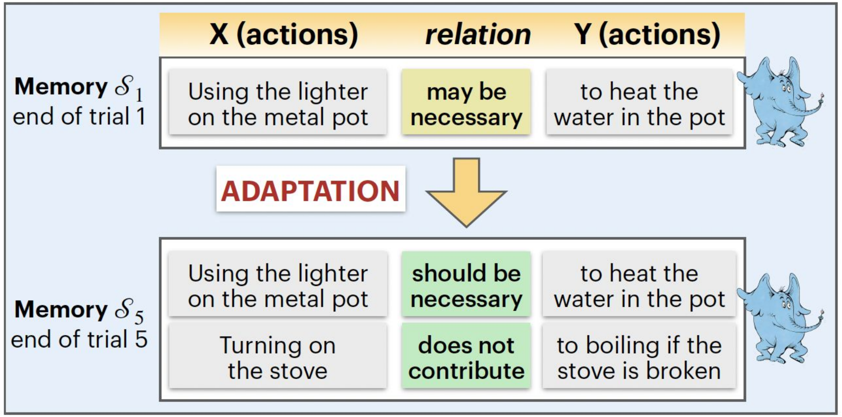
Learning **state transitions** is essential for SDM

1. actions enabling **desired** state transitions
2. actions producing **undesired** or no changes
3. state transitions **contributing to the task**

A collection of natural language statements capturing **causal abstractions of action-effects** *favorable to exploit at test-time like hindsight experience replay*

Good effects: X → is necessary to → Y
Bad effects: X → does not contribute → Y

Uncertainty
Low: may be; **High:** should be



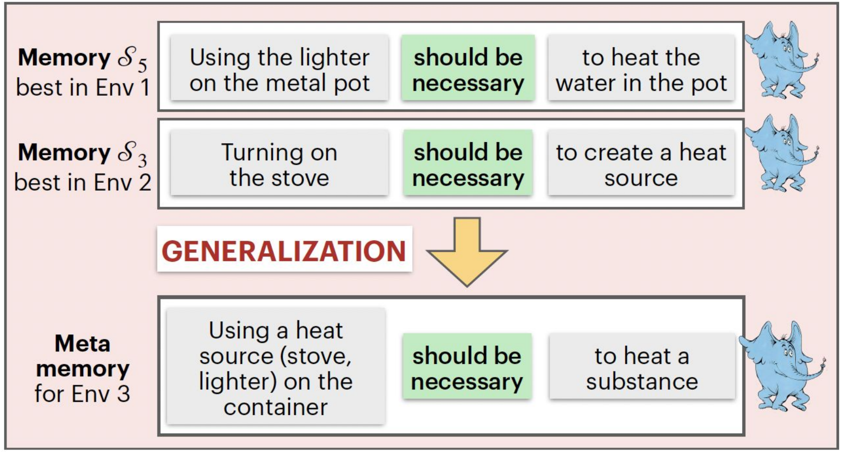
Meta-Memory

Task- and environment-specific memory cannot help generalize such as knowing how to <u>boil water</u> may not help knowing how to <u>boil cadmium</u> unless ***generalized abstractions***.

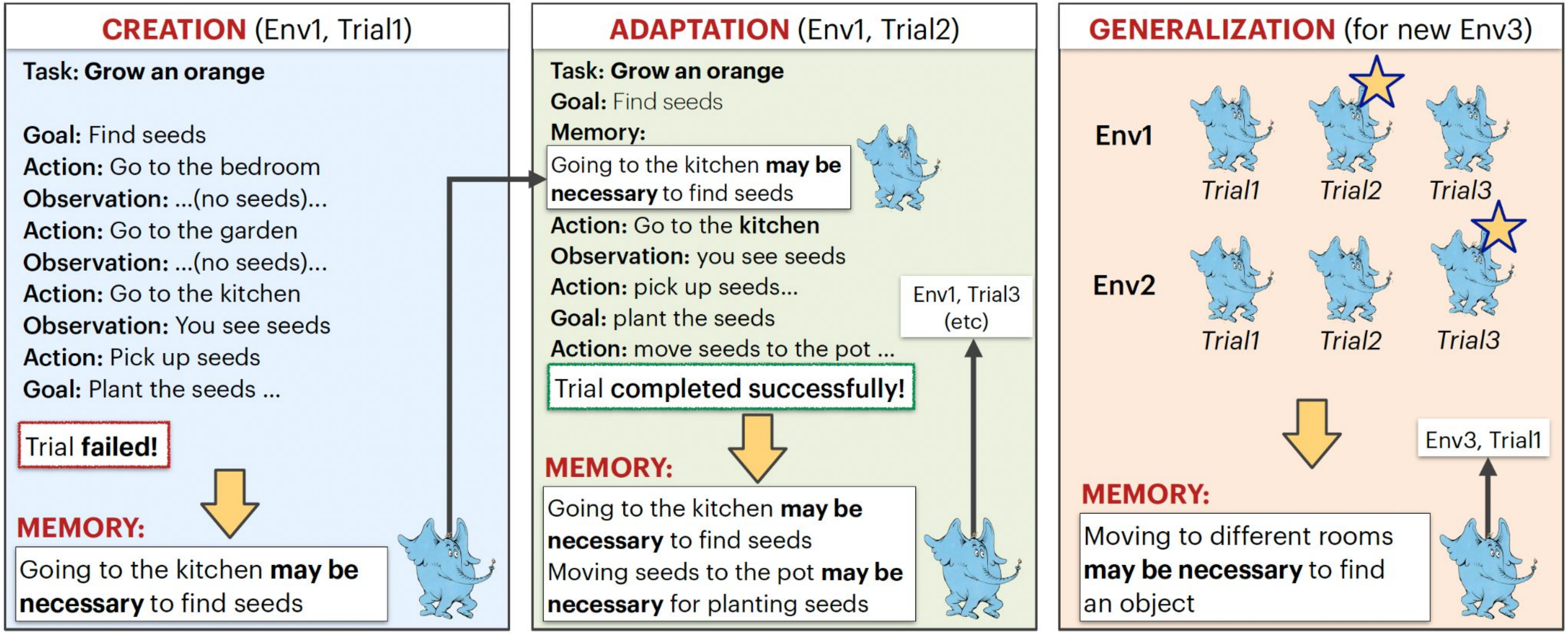
Select the best memories from past attempts across diverse environments/tasks
auto-curriculum selection

Meta-memory with *generalized* instruction:

“Generate insights to solve the same task in a new environment configuration”



CLIN: Summary



Outline

Background

CLIN: Continual Learning from Interactions

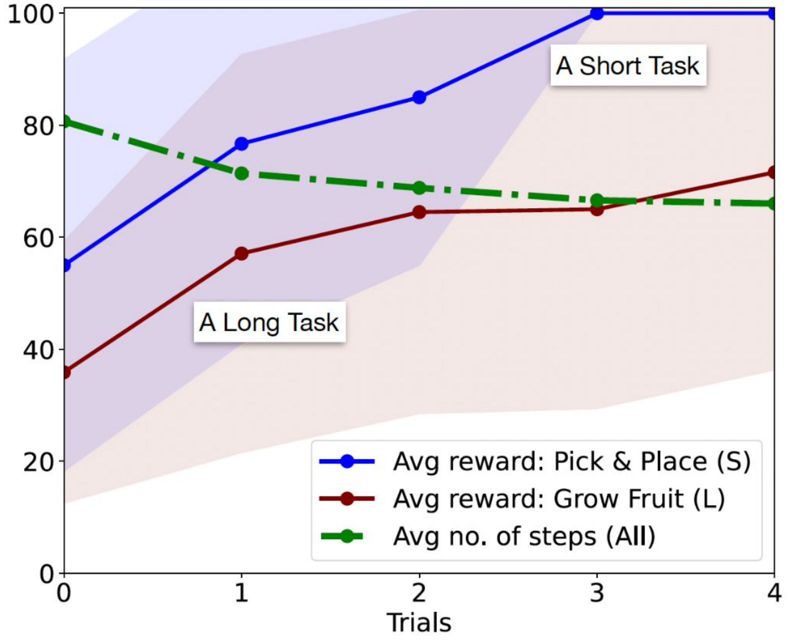
- Proposed Architecture
- What does CLIN learn over time?
- **Results on ScienceWorld & ALFWorld**

SSO: Skill Set Optimization

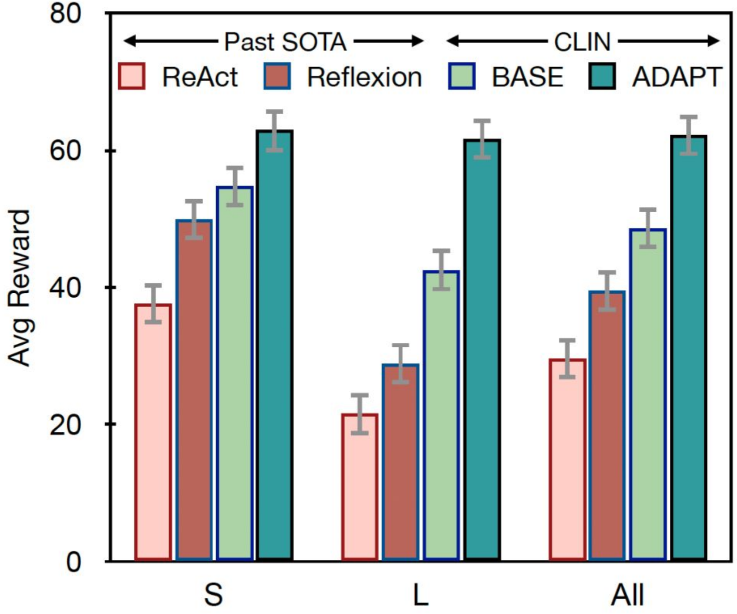
- Skills
- Skill Set Optimization
- Results on ScienceWorld & NetHack

CLIN Exhibits Rapid Task Adaptation

Quick adaptation, improved efficiency



CLIN beats reflective SOTA



CLIN Generalizes to Novel Environments

Train:
Boil water
Boil chocolate

Test:
Boil Cadmium

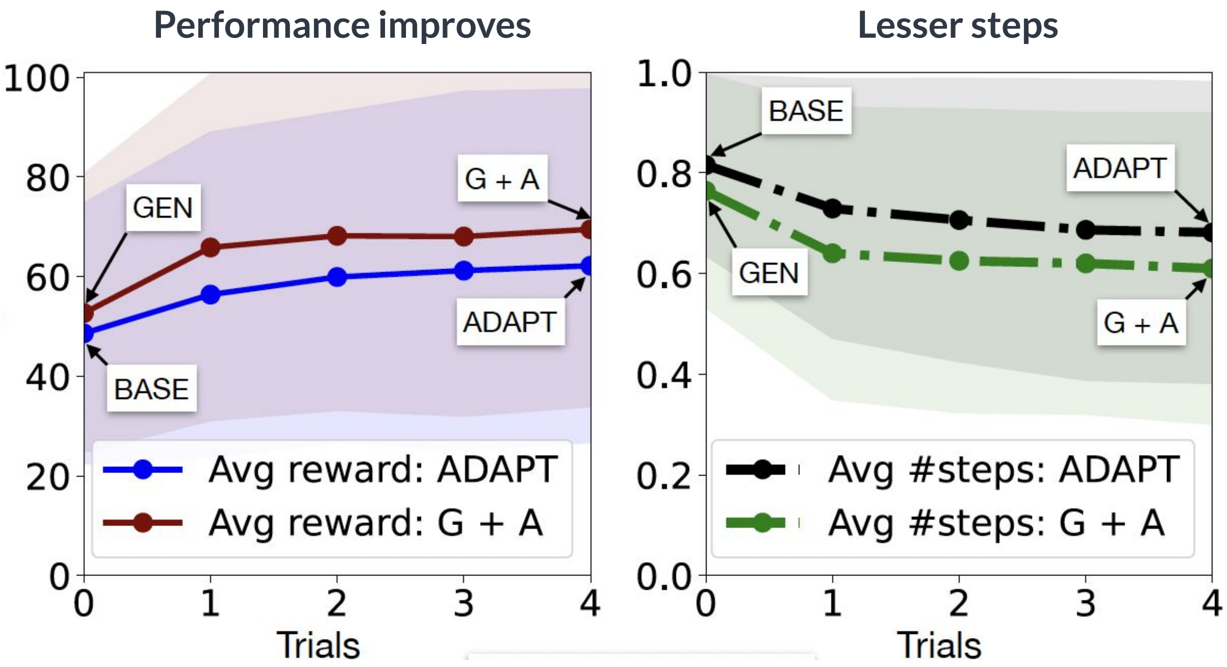
CLIN even beats imitation learning baselines (that uses gold trajectories) in most lengthy, complex tasks

		RL Methods			Generative Language Agents			CLIN (ours)		
Task	**Type**	**DRRN**	**KGA2C**	**CALM**	**SayCan**	**ReAct**	**Reflexion**	**BASE**	**GEN-ENV**	**G+A**
Temp_1	S	6.6	6.0	1.0	**26.4**	7.2	5.9	25.2	15.7	13.8
Temp_2	S	5.5	11.0	1.0	8.0	6.1	28.6	53.2	49.7	**58.2**
$\text{Pick\&Place}_1$	S	15.0	18.0	10.0	22.9	26.7	64.9	92.5	59.2	**100.0**
$\text{Pick\&Place}_2$	S	21.7	16.0	10.0	20.9	53.3	16.4	55.0	**100.0**	**100.0**
Chemistry_1	S	15.8	17.0	3.0	47.8	51.0	**70.4**	44.5	42.2	51.7
Chemistry_2	S	26.7	19.0	6.0	39.3	58.9	70.7	56.7	85.6	**93.3**
Lifespan_1	S	50.0	43.0	6.0	80.0	60.0	**100.0**	85.0	65.0	**100.0**
Lifespan_2	S	50.0	32.0	10.0	67.5	67.5	84.4	70.0	75.0	**90.0**
Biology_1	S	8.0	10.0	0.0	16.0	8.0	8.0	10.0	32.0	**32.0**
Boil	L	3.5	0.0	0.0	**33.1**	3.5	4.2	7.0	4.4	16.3
Freeze	L	0.0	4.0	0.0	3.9	7.8	7.8	**10.0**	8.9	**10.0**
GrowPlant	L	8.0	6.0	2.0	9.9	9.1	7.3	10.2	10.9	**11.2**
GrowFruit	L	14.3	11.0	4.0	13.9	18.6	13.0	35.9	70.8	**94.5**
Biology_2	L	21.0	5.0	4.0	20.9	27.7	2.6	70.0	42.8	**85.6**
Force	L	10.0	4.0	0.0	21.9	40.5	50.6	53.5	70.0	**100.0**
Friction	L	10.0	4.0	3.0	32.3	44.0	**100.0**	56.5	70.0	94.0
Genetics_1	L	16.8	11.0	2.0	67.5	25.7	50.9	77.4	84.5	**100.0**
Genetics_2	L	17.0	11.0	2.0	59.5	16.8	23.7	62.3	61.4	**100.0**
	S	22.1	19.1	5.2	36.5	37.6	49.9	54.7	58.3	**71.0**
	L	11.2	6.2	1.9	29.2	21.5	28.9	42.5	47.1	**68.0**
	All	16.7	12.7	3.6	32.9	29.6	39.4	48.6	52.7	**69.5**

Efficient Generalization

Performance drops 6.2 point and in 10% episodes if we do not use **causal format** for memory insights

Controller adds 18 points to a base (ReAct) performance improving 44% episodes

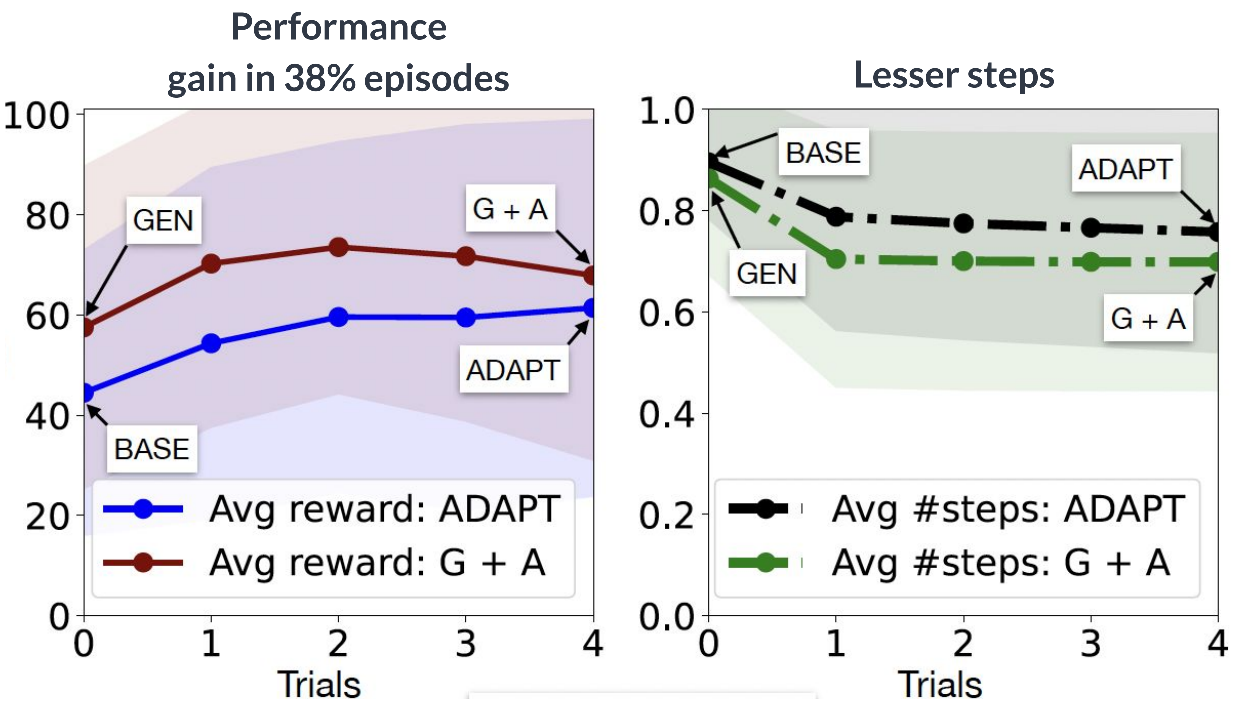


CLIN Generalizes to Novel Tasks

Train (in Env 1):
Boil water
Boil apple juice

Test (in Env 1):
Freeze Water

The improvement attributes to ***critical learning about the environment***
(apple juice is in the fridge)



Memory Precision

Natural selection of good memory items over time shows CLIN can auto-correct when the starting memory is not applicable due to loss of specificity or lack of information.

CLIN converges to a more precise representation of the world

	GEN-ENV (Trial 0)	GEN-ADAPT (Best Trial)
No. of insights	100	105
Correct insights	72.0%	**91.4%**
Final score	39.1	**55.9**

	GEN-TASK (Trial 0)	GEN-ADAPT (Best Trial)
No. of insights	98	107
Correct insights	73.9%	**91.1%**
Final score	43.7	**58.1**

Is Causal Abstraction Helpful?

Memory with no structure is generic ("Be clear with your actions"), often contains ungrounded information ("use a food processor"), and does not naturally abstract causal relations towards a world model ("this is unnecessary and wastes time")

Ablations for CLIN

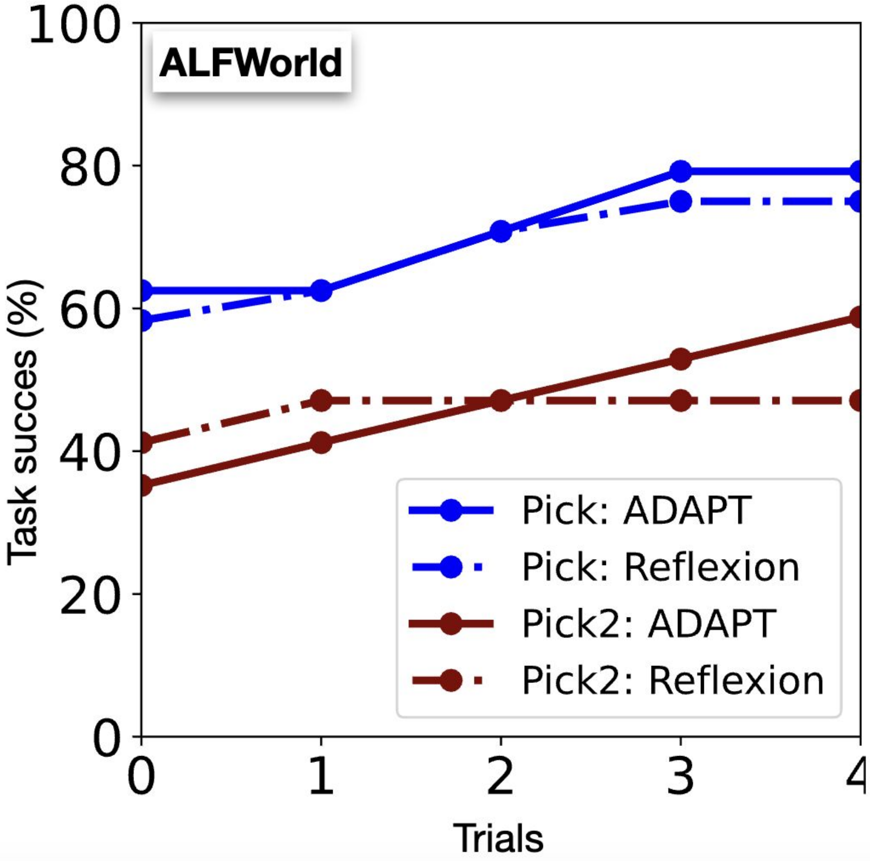
Ablation Setup	**Δavg score (↓)**	**% ep. drop. (↑)**
Abl-Causal-Memory	-6.2	10
Abl-Controller-BASE	-18.1	44.8

ALFWorld

You are in the middle of a room. Looking quickly around you, you see a drawer 2, a shelf 5, a drawer 1, a shelf 4, a sidetable 1, a drawer 5, a shelf 6, a shelf 1, a shelf 9, a cabinet 2, a sofa 1, a cabinet 1, a shelf 3, a cabinet 3, a drawer 3, a shelf 11, a shelf 2, a shelf 10, a dresser 1, a shelf 12, a garbagecan 1, a armchair 1, a cabinet 4, a shelf 7, a shelf 8, a safe 1, and a drawer 4.

Your task is to: *put some vase in safe.*

> go to shelf 6
You arrive at loc 4. On the shelf 6, you see a vase 2.



Failures

CLIN is able to **compose insights**

No stove, use furnace (Env 1) + Go to Kitchen for apple juice (Env 2)

But when it **fails**, it is due to:

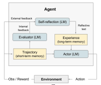
1. **Lack of exploration**

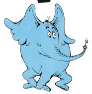
   If it has never visited an art studio,
   it will never "explore" to reach art studio for collecting paints

2. **Poor memory retrieval**

   It knows to use stove for heating OR use furnace when stove is broken
   BUT to boil cadmium it needs to use furnace even if the stove is working

Failures



	Transferable		Scalable	
	Multi-task	Modular	Lossless	Sublinear
Fewshot	X	X	✓	X
Reflexion	X	X	X	X
ExpeL	✓	X	✓	✓
Voyager	X	✓	✓	X
CLIN	✓	✓	X	✓

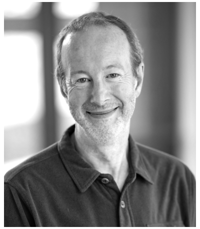
1. Shinn, Noah, Federico Cassano, Beck Labash, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. "Reflexion: Language agents with verbal reinforcement learning." arXiv preprint arXiv:2303.11366 (2023).
2. Zhao, Andrew, Daniel Huang, Quentin Xu, Matthieu Lin, Yong-Jin Liu, and Gao Huang. "ExpeL: LLM Agents Are Experiential Learners." arXiv preprint arXiv:2308.10144 (2023).
3. Wang, Guanzhi, Yuqi Xie, Yunfan Jiang, Ajay Mandlekar, Chaowei Xiao, Yuke Zhu, Linxi Fan, and Anima Anandkumar. "Voyager: An open-ended embodied agent with large language models." arXiv preprint arXiv:2305.16291 (2023).
4. Majumder, Bodhisattwa Prasad, Bhavana Dalvi Mishra, Peter Jansen, Oyvind Tafjord, Niket Tandon, Li Zhang, Chris Callison-Burch, and Peter Clark. "CLIN: A Continually Learning Language Agent for Rapid Task Adaptation and Generalization." arXiv preprint arXiv:2310.10134 (2023).

SSO:
Skill Set Optimization

Kolby, **Bodhi**, Bhavana,
Sameer, Pete, Roy








ARISTO

UCI

Ai2

Outline

Background

CLIN: Continual Learning from Interactions

- Proposed Architecture
- What does CLIN learn over time?
- Results on ScienceWorld & ALFWorld

SSO: Skill Set Optimization

- **Skills**
- Skill Set Optimization
- Results on ScienceWorld & NetHack

Skills

- World model information should:
  - Be general, composable, editable, and retrievable
  - Contribute to LLM agent's knowledge of the world model (state & action transitions)

Skill Definition

Target:

- goal state feature

Prerequisites:

- initial state features
- used for retrieval

Instructions:

- generic actions to execute

Example generated Skill

Target: agent is in the 'target location'

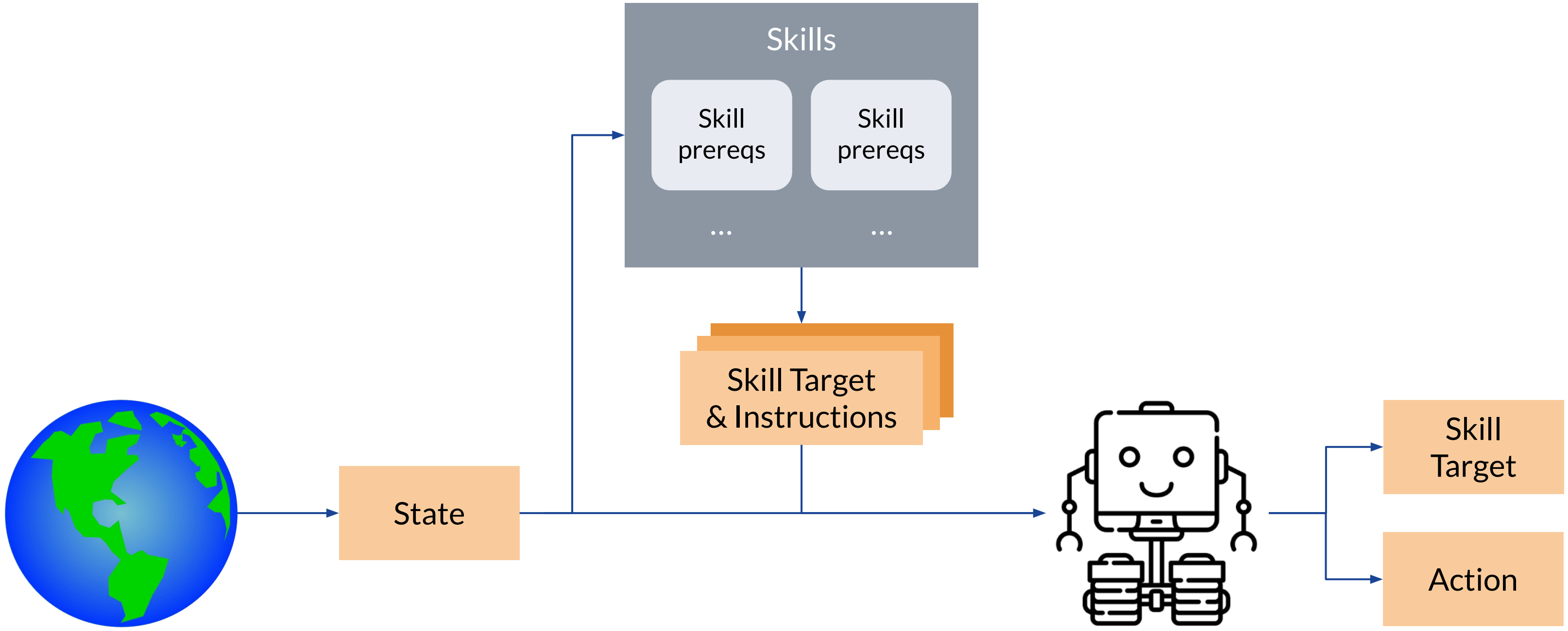
Prereqs:

1. agent is in a location that has a door leading to a hallway
2. there exists a known target location to which agent needs to move
3. agent is able to move (not restricted or blocked)

Instructions:

1. go to hallway
2. go to 'target location'

Using Skills



Outline

Background

CLIN: Continual Learning from Interactions

- Proposed Architecture
- What does CLIN learn over time?
- Results on ScienceWorld & ALFWorld

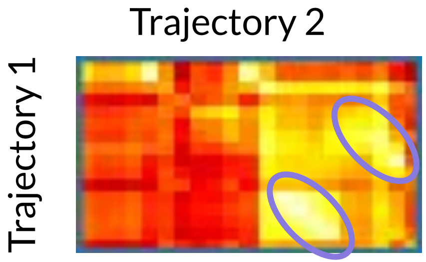
SSO: Skill Set Optimization

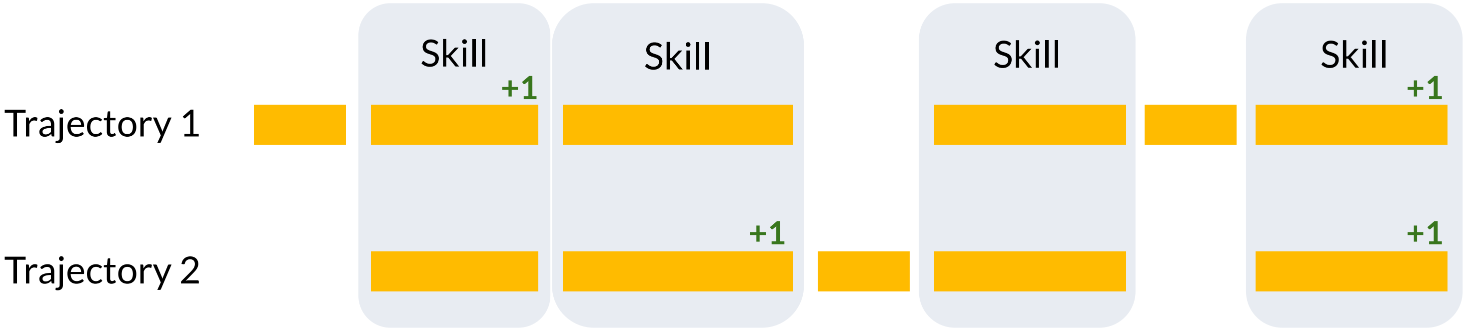
- Skills
- **Skill Set Optimization**
- Results on ScienceWorld & NetHack

Skill Set Optimization

We want to extract reusable sequences that lead to rewards

1. Find common sub trajectories
   a. Trim trajectories to end in positive rewards
   b. Align sub trajectories using state, action embedding from LLM





Skill Set Optimization

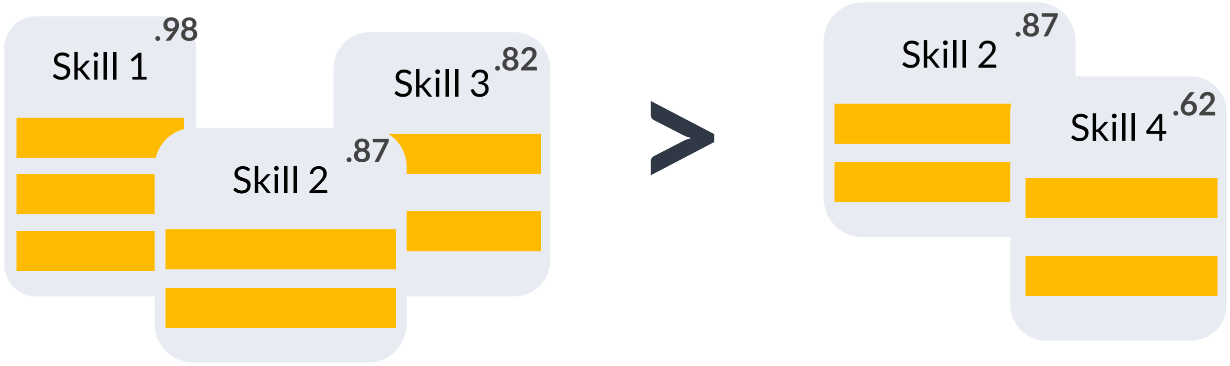
We want to extract reusable sequences that lead to rewards

1. Find common sub trajectories
2. Score and sort skills
    a. Similarity
    b. Reward
    c. Length

Skill Set Optimization

We want to extract reusable sequences that lead to rewards

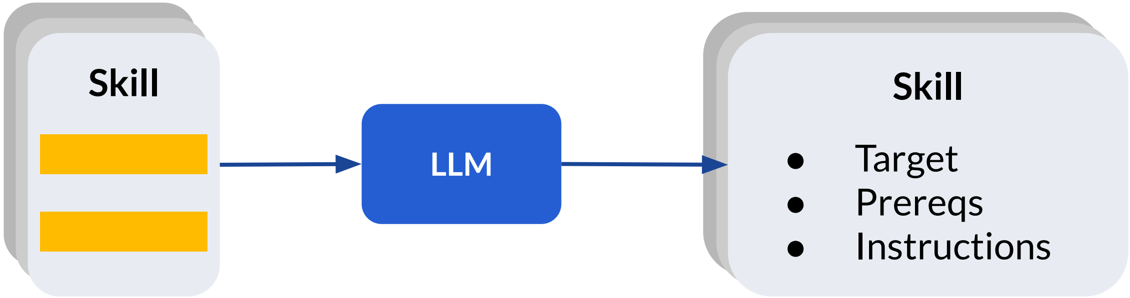
1. Find common sub trajectories
2. Score and sort skills
3. Construct skill set using beam search
   a. Do not allow skill sub trajectories to overlap
   b. Select best beam based on sum of scores: similarity, reward, length



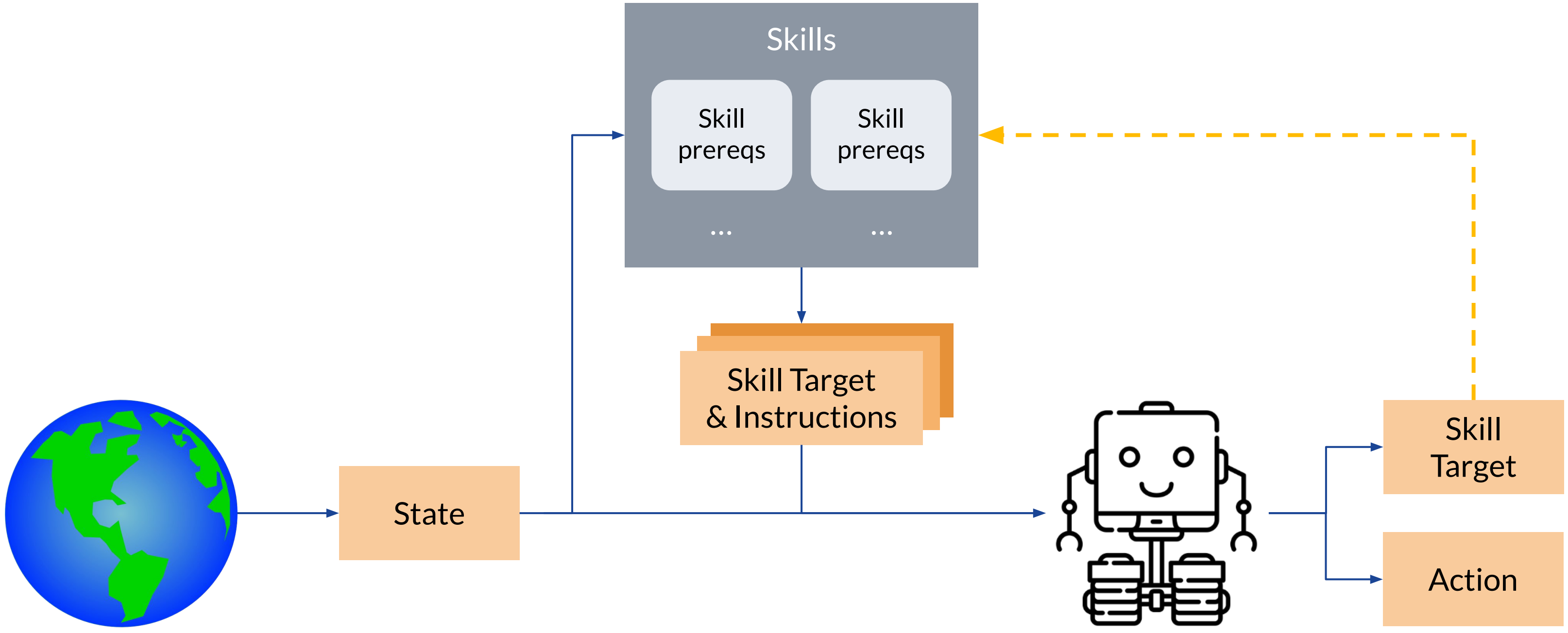
Skill Set Optimization

We want to extract reusable sequences that lead to rewards

1. Find common sub trajectories
2. Score and sort skills
3. Construct skill set using beam search
4. Generate skill target, prerequisites, and instructions

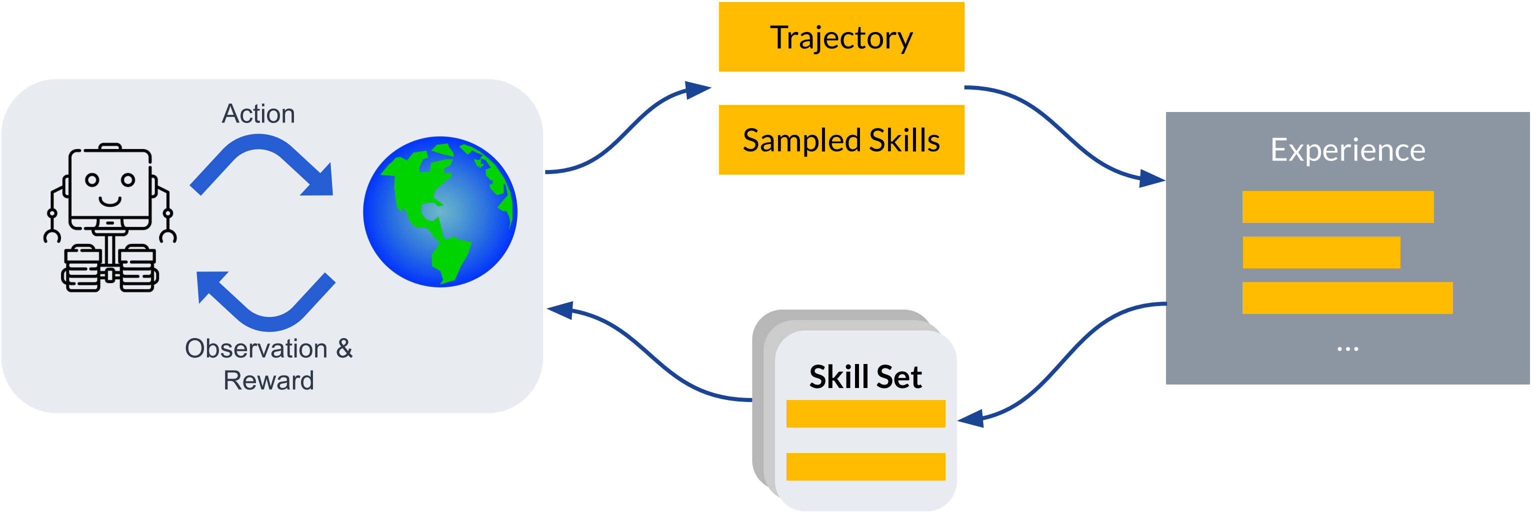


Skill Set Optimization



Skill Set Optimization

- Prioritize sampled skills that lead to positive reward
- Black list sampled skills that lead to negative reward
- After every trajectory, extract skills from last N trajectories



Outline

Background

CLIN: Continual Learning from Interactions

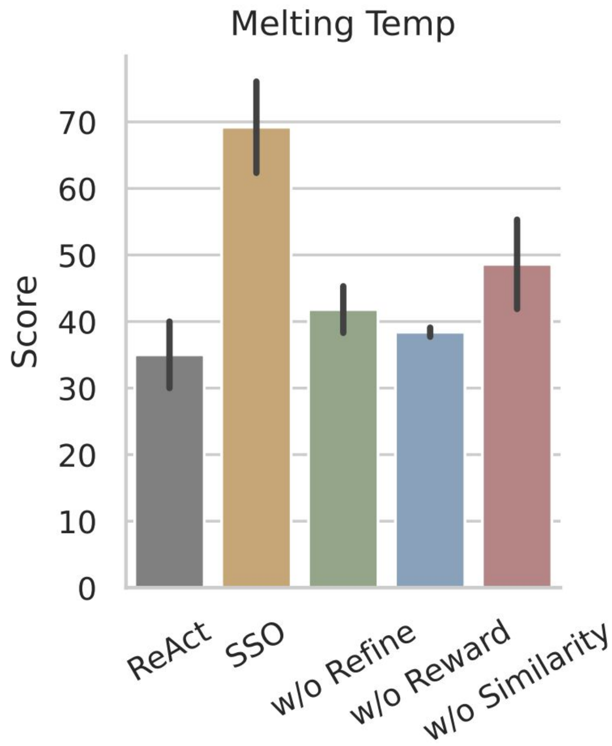
- Proposed Architecture
- What does CLIN learn over time?
- Results on ScienceWorld & ALFWorld

SSO: Skill Set Optimization

- Skills
- Skill Set Optimization
- **Results on ScienceWorld & NetHack**

SSO improves over CLIN

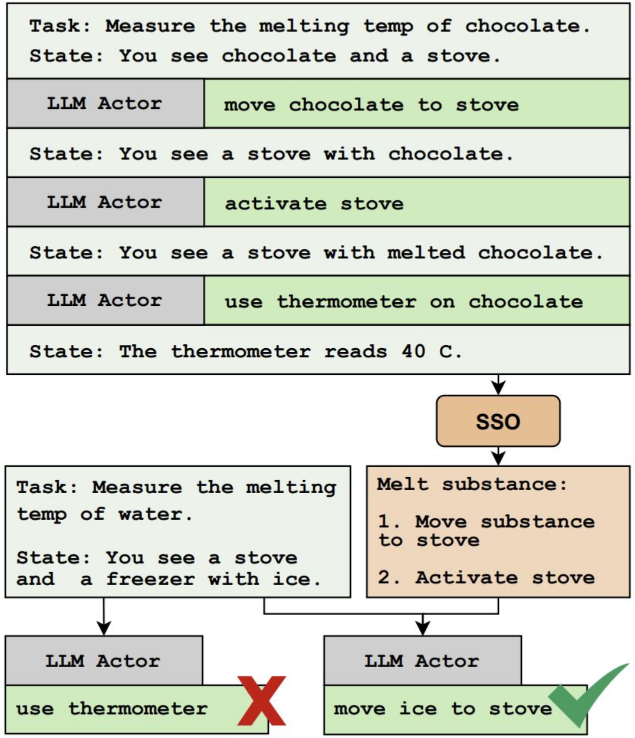
ScienceWorld Task	ReAct	Adaptation Reflexion	CLIN	SSO	Transfer CLIN	SSO
Temperature	7.2	5.9	14.3	**100**	15.7	**71.6**
Melting Temp	6.1	28.6	51.8	**97.3**	49.7	**69.2**
Find Plant	26.7	64.9	**100**	**100**	59.2	**100**
Find Living	53.3	16.4	**100**	96.7	**100**	90
Chemistry	51	70.4	44.4	**82.6**	42.2	**48**
Color Mixing	58.9	70.7	56.7	**81.1**	**85.6**	71.1
Lifespan, Longest	61	**100**	**100**	**100**	65	**90**
Lifespan, Shortest	67.5	84.4	90	**100**	75	**80**
Life Stages, Plant	8	**8**	**8**	6.2	**32**	3.4
Life Stages, Animal	27.7	2.6	81	**100**	42.8	**77**
Boil	3.5	4.2	15.2	**81.7**	4.4	**48.7**
Freeze	7.8	7.8	10	**74.3**	8.9	**38.9**
Grow Plant	9.1	7.3	11	**86.6**	10.9	**61.2**
Grow Fruit	18.6	13	71.6	**78**	**70.8**	28.3
Gravity	40.5	50.6	**100**	**100**	70	**74**
Friction	44	**100**	72.5	94	**70**	67.5
Genetics, Known	25.7	50.9	**100**	78.5	**84.5**	42.5
Genetics, Unknown	16.8	23.7	**92.6**	48.7	**61.4**	20.3
Average	29.6	39.4	62.2	**83.7**	52.7	**60.1**



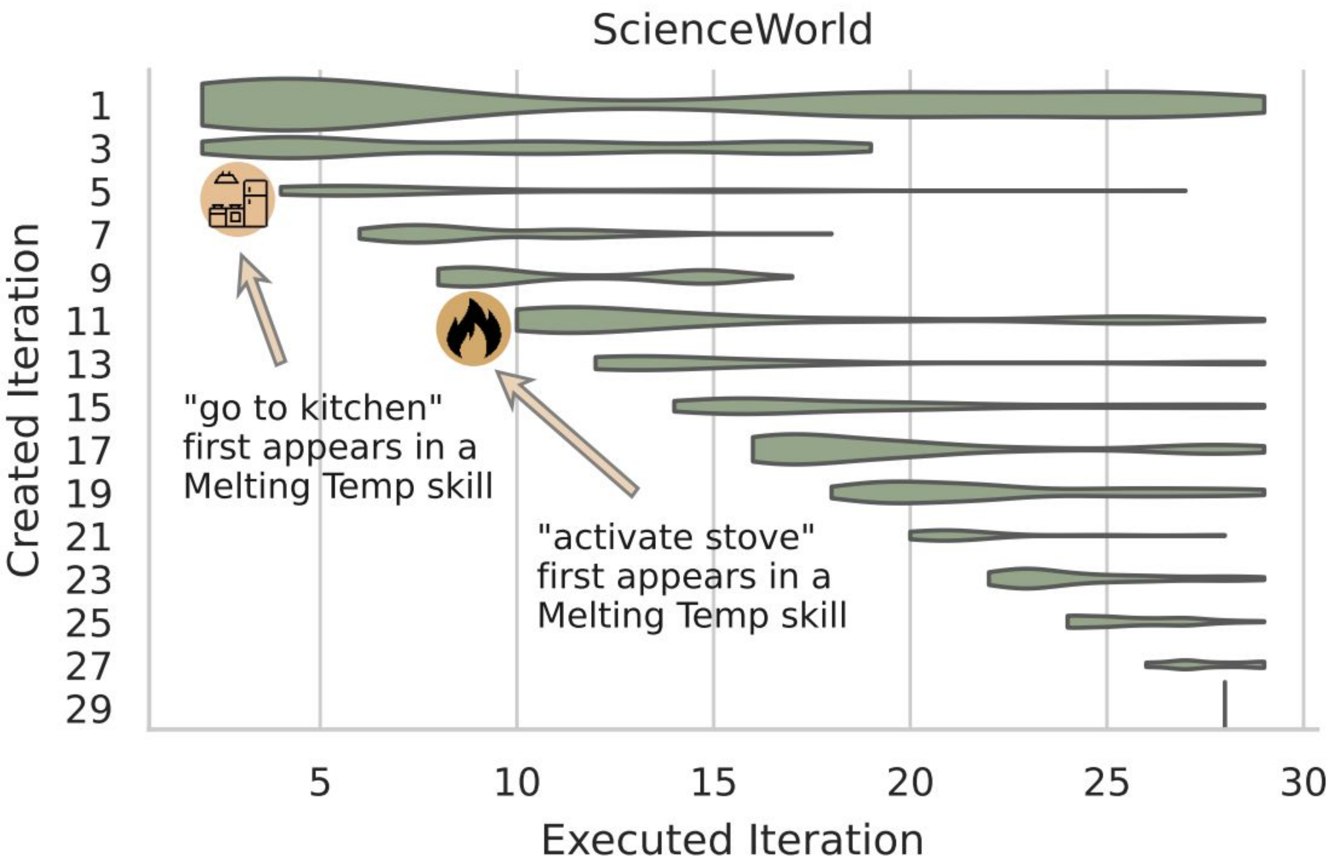
An example skill

ScienceWorld Melting Temp Task

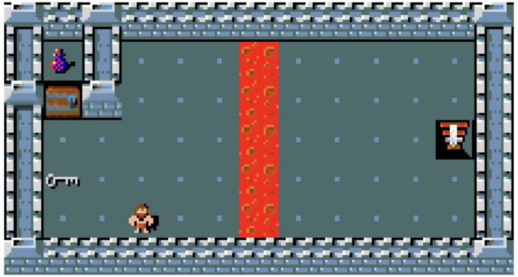
Subgoal: The stove is turned on. on the stove is: a substance called liquid [substance].
1. Focus on the thermometer
2. Focus on the substance you want to heat
3. Move the focused substance to the stove
4. Activate the stove

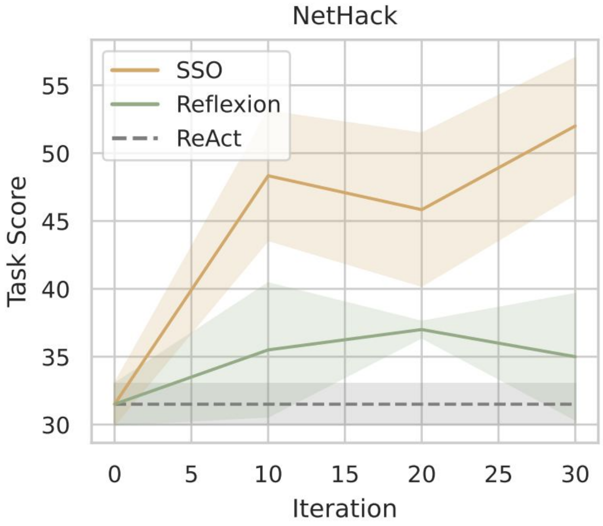


Skill Lifecycle

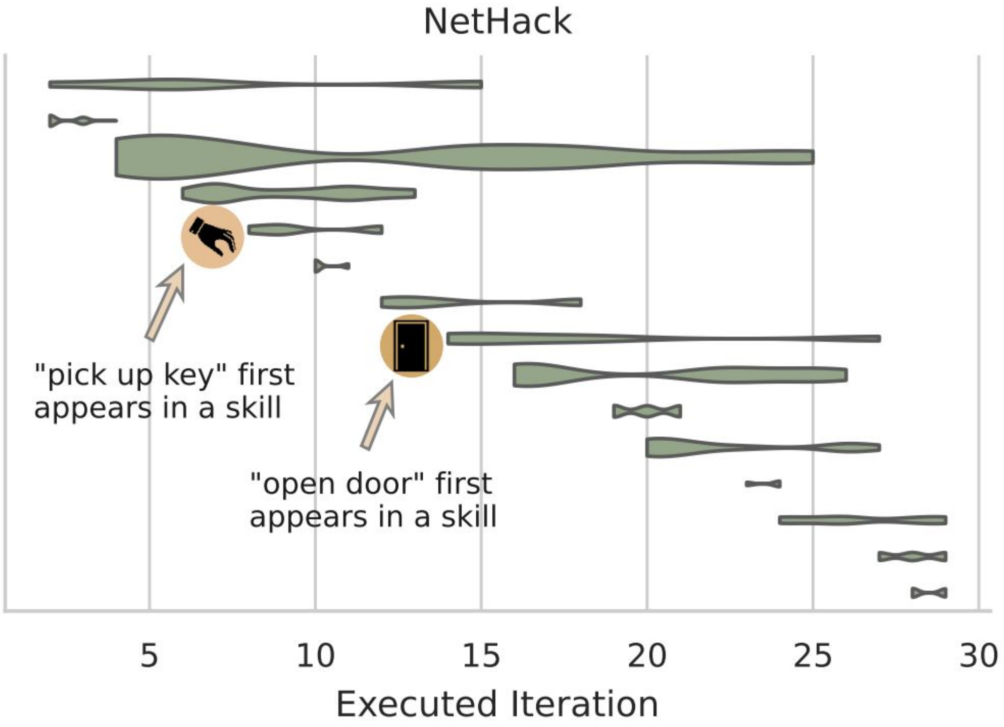


NetHack





NetHack
SSO
Reflexion
ReAct
Task Score
55
50
45
40
35
30
0
5
10
15
20
25
30
Iteration



NetHack
"pick up key" first appears in a skill
"open door" first appears in a skill
5
10
15
20
25
30
Executed Iteration

Conclusion

CLIN: https://allenai.github.io/clin/

CLIN: A CONTINUALLY LEARNING LANGUAGE AGENT FOR RAPID TASK ADAPTATION AND GENERALIZATION

Bodhisattwa Prasad Majumder[1], **Bhavana Dalvi Mishra**[1], **Peter Jansen**[1,2], **Oyvind Tafjord**[1], **Niket Tandon**[1], **Li Zhang**[3], **Chris-Callison Burch**[3], **Peter Clark**[1]
[1]Allen Institute of AI
[2]University of Arizona
[3]University of Pennsylvania

SSO: https://allenai.github.io/sso/

Skill Set Optimization:
Reinforcing Language Model Behavior via Transferable Skills

Kolby Nottingham[1] **Bodhisattwa Prasad Majumder**[*2] **Bhavana Dalvi**[*2]
Sameer Singh[1] **Peter Clark**[2] **Roy Fox**[1]

- Dynamic memory in the form causal abstractions or skills helps in generalization
- But current execution is greedy best, can we improve?
- Memory helps exploiting world knowledge, but how to incentivize exploration?

Thank you!