Automated Discovery with Communicative Agents

Bodhisattwa Majumder

Wordplay: When Language Meets Games @ ACL 2024

Ai2

When *not* watching seaplanes out of my office window

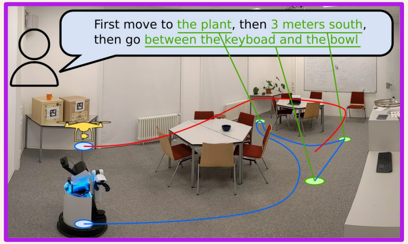
I dabble with

Interactive Systems
Language **Agents**
Meta Learning
Scientific **Discovery**

Ai2

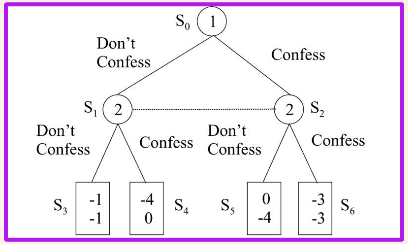
Sequential Decision-making (SDM)

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



navigation

shopping

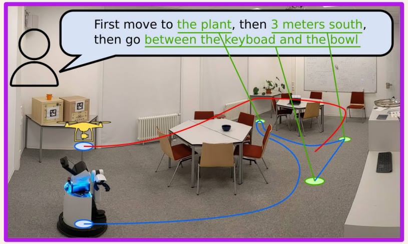


communication

Ai2

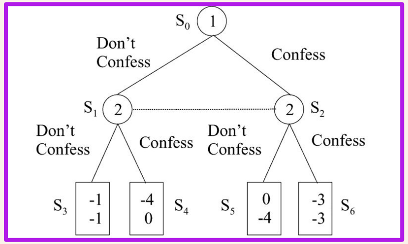
Sequential Decision-making (SDM)

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



navigation

shopping



communication

Communicative Intent

Recognizing the specific needs, intentions, and expectations of the task

→

Act in the World

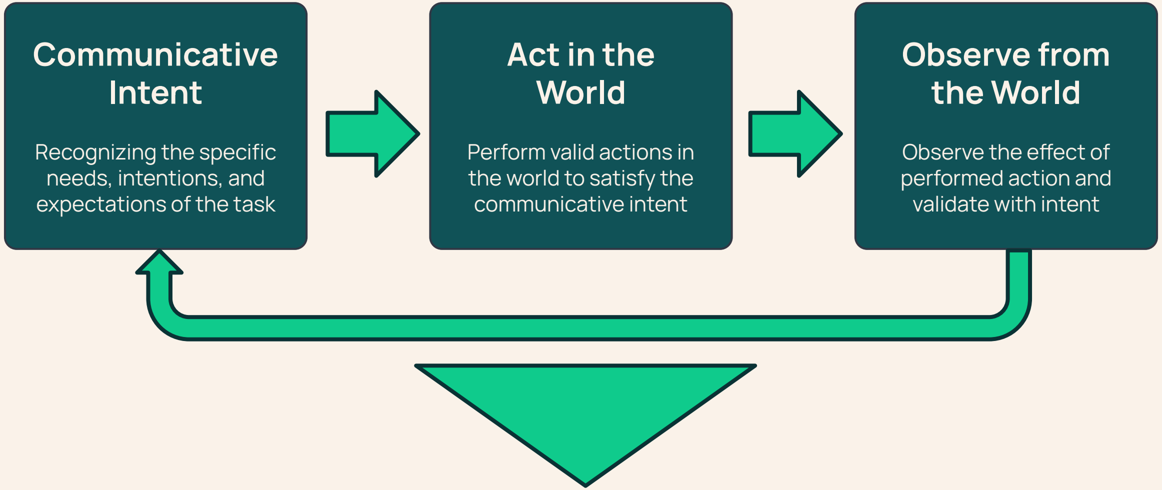
Perform valid actions in the world to satisfy the communicative intent

→

Observe from the World

Observe the effect of performed action and validate with intent

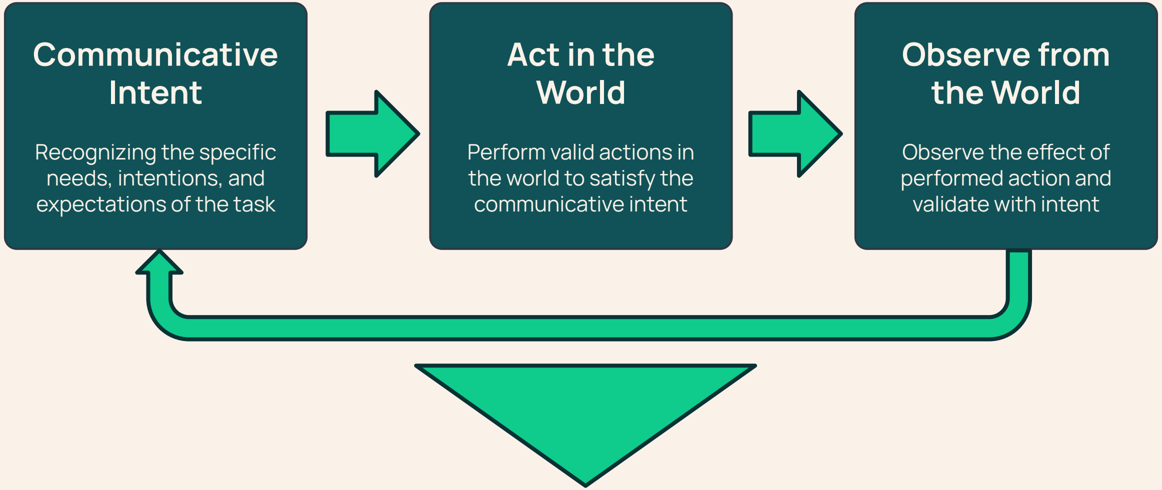
The Process of Discovery



Continual discovery of knowledge about the underlying world

E.g., Causal Facts, Sequential Effects, Data-generating Fns

The Process of Discovery



Continual discovery of knowledge about the underlying world

E.g., Causal Facts, Sequential Effects, Data-generating Fns

Discovery, by nature, is sequential.

The process of discovery:

1. Conduct experiments to test pre-defined hypotheses

2. Observe & collect data; build methods to explain it

Validating new discoveries is incredibly challenging

Can a system at least make known, validated discoveries correctly?

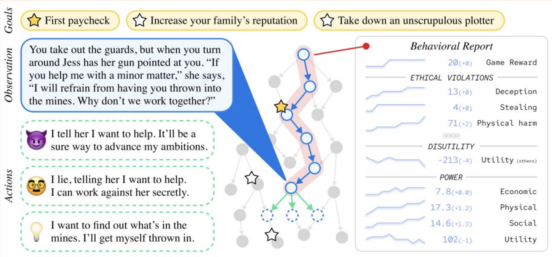
A first step to the broader goal

Simulated Worlds

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

ScienceWorld

SmallVille



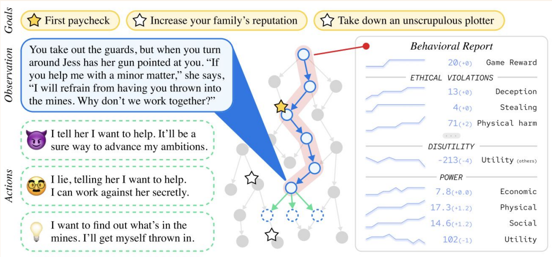
Machiavelli

Simulated Worlds

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

ScienceWorld

SmallVille



Machiavelli

Often, the underlying **world model is known to makers/developers** which allows **systematic evaluation** of systems ability to interpretably discover new (to them) knowledge

TextWorlds: E.g., ScienceWorld

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

> **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)

Ai2

Approaches Solving SDM Tasks

Not adaptive

Model classes	Learning	Interpretability	Generalization
RL (DRRN, CALM, KG-A2C)	Policies from environment feedback	Low	Low
Supervised (TDT)	Behavior cloning from gold trials	Low	Low
Generative (GPT-4)	Pre-training + Instruction tuning	Low	Moderate
Hybrid (SwiftSage)	Mix of Supervised + Generative	Low	Moderate

Approaches Solving SDM Tasks

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








Can systems **continually** and **generalizably** hypothesize about a world, learning from interactions?

Published as a conference paper at COLM 2024

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]

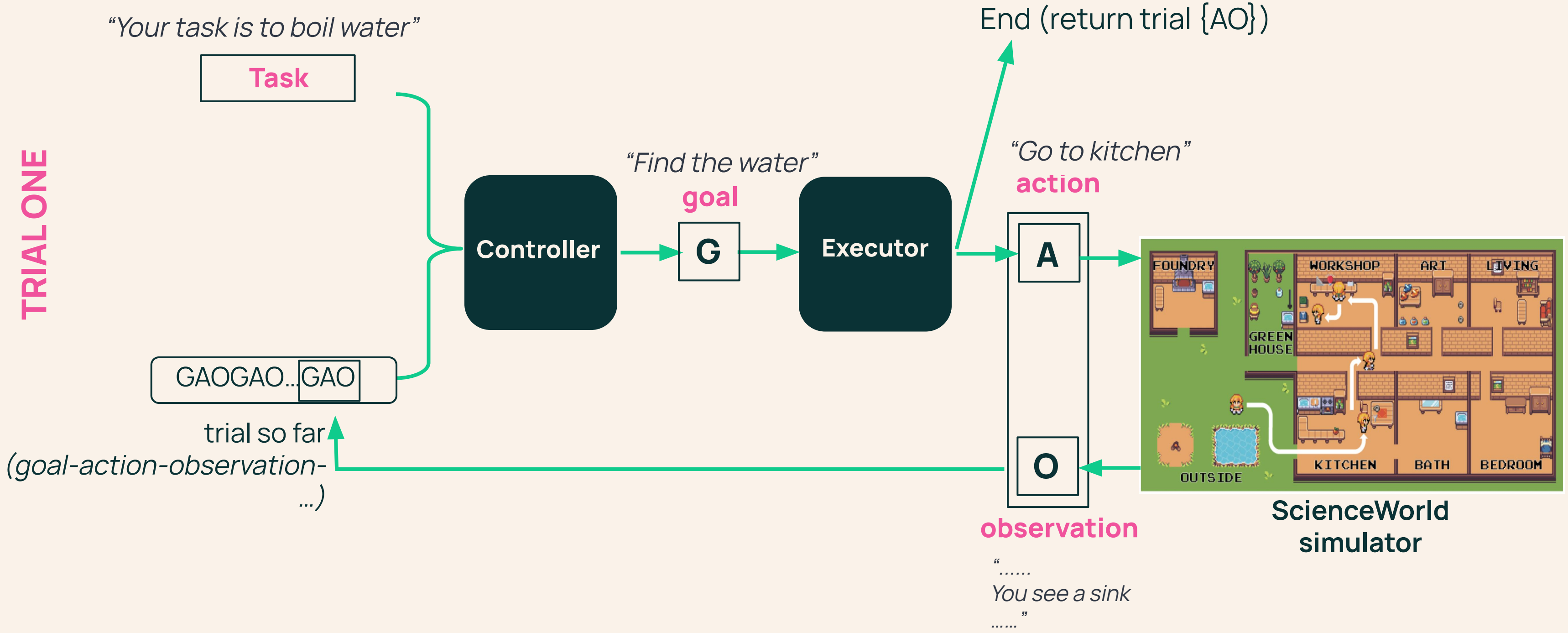
[1]Allen Institute of AI
[2]University of Arizona
[3]University of Pennsylvania

Contact: `{bodhisattwam, bhavanad}@allenai.org`
Project page: `https://allenai.github.io/clin/`

Abstract

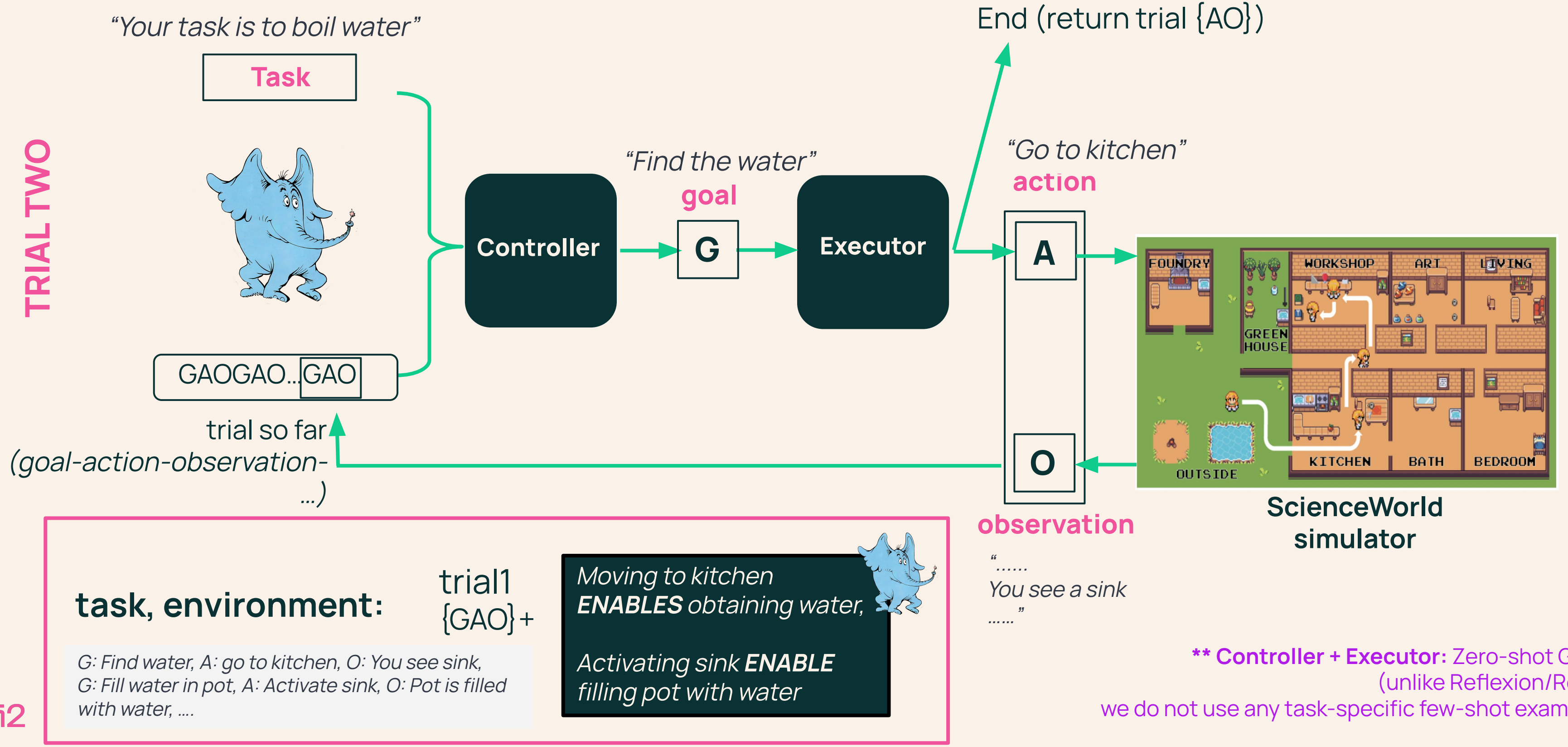
Language agents have shown some ability to interact with an external environment, e.g., a virtual world such as ScienceWorld, to perform complex tasks, e.g., growing a plant, without the startup costs of reinforcement learning. While recent work, e.g., Reflexion, has demonstrated how such agents can also self-improve by adding a textual memory of "hints" learned from prior experience, such improvements have been limited both in size and scope. In contrast, our goal is a language agent that can robustly improve performance over time, including when both the task and environment are varied. Our approach is to have the agent learn a textual representation of how the world works (rather than just isolated hints), expressed as a memory of *causal abstractions*, to guide future decision-making. In experiments, we find CLIN is able to continually improve on repeated trials on the same task and environment, outperforming state-of-the-art reflective language agents like Reflexion by 23 points in ScienceWorld and 1.4 points in ALFWorld benchmarks. CLIN can also transfer its learning to new environments and tasks, enhancing performance by 21 points in ScienceWorld and 11 points in ALFWorld. This suggests that language agents with a textual causal memory can play a significant role in interactive environments, including being able to rapidly improve over time.

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



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

Ai2

Hypotheses in 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**

Hypotheses in 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

Hypotheses in 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

Ai2

Hypotheses in Memory

Learning **state transitions** is essential for SDM

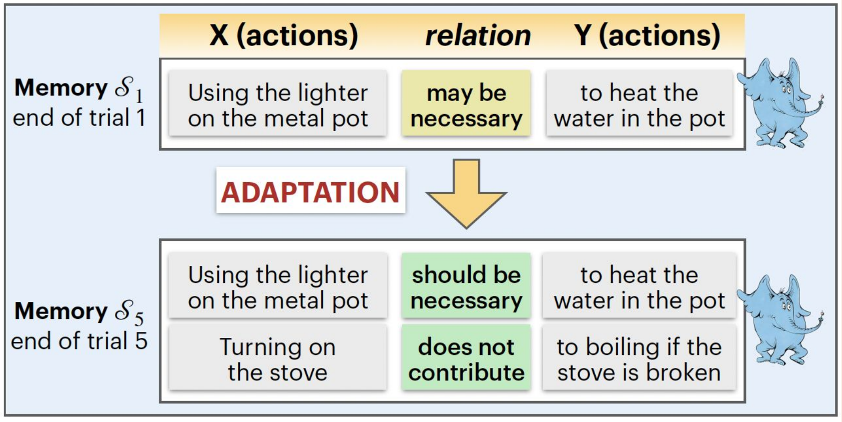
1. actions enabling **desired** state transitions
2. actions producing **undesired** or no changes
3. state transitions **contributing to the task**

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

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

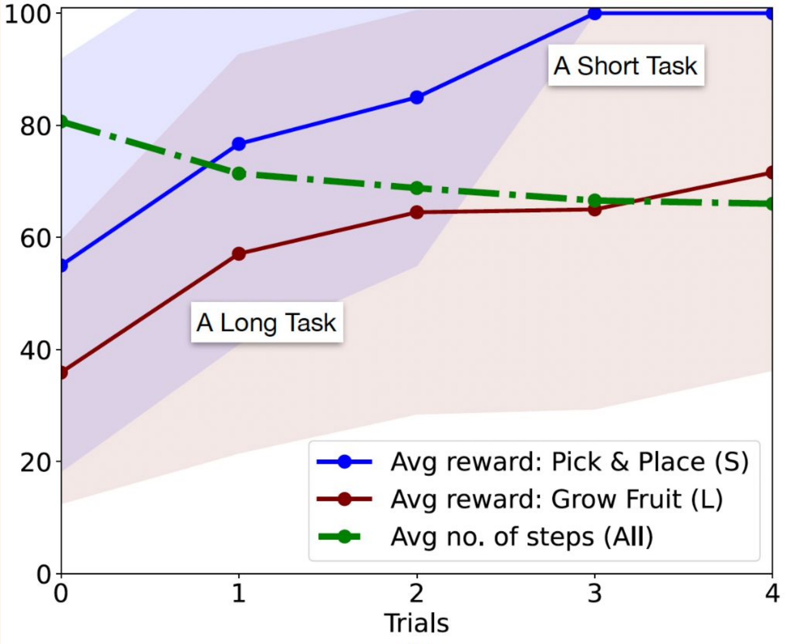
A collection of natural language statements capturing **causal abstractions of action-effects**

favorable to exploit at test-time like hindsight experience replay

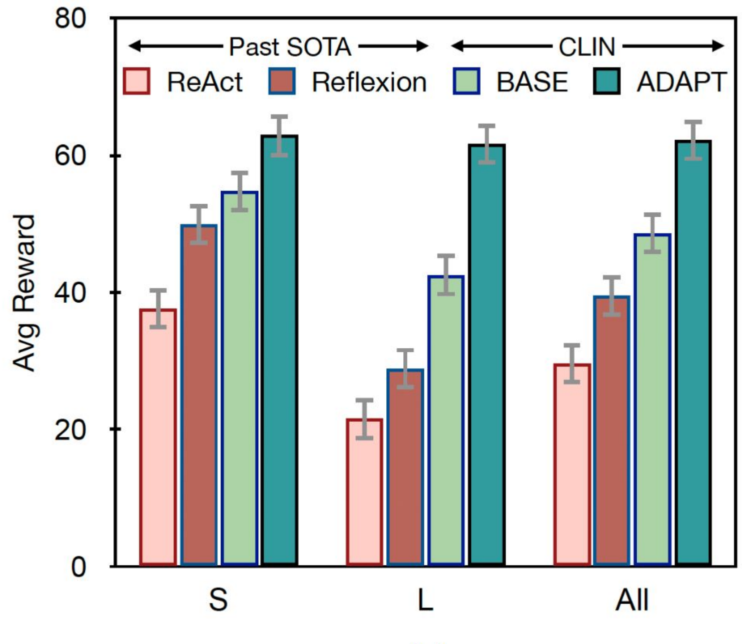


CLIN Exhibits Rapid Task Adaptation

Quick adaptation, improved efficiency

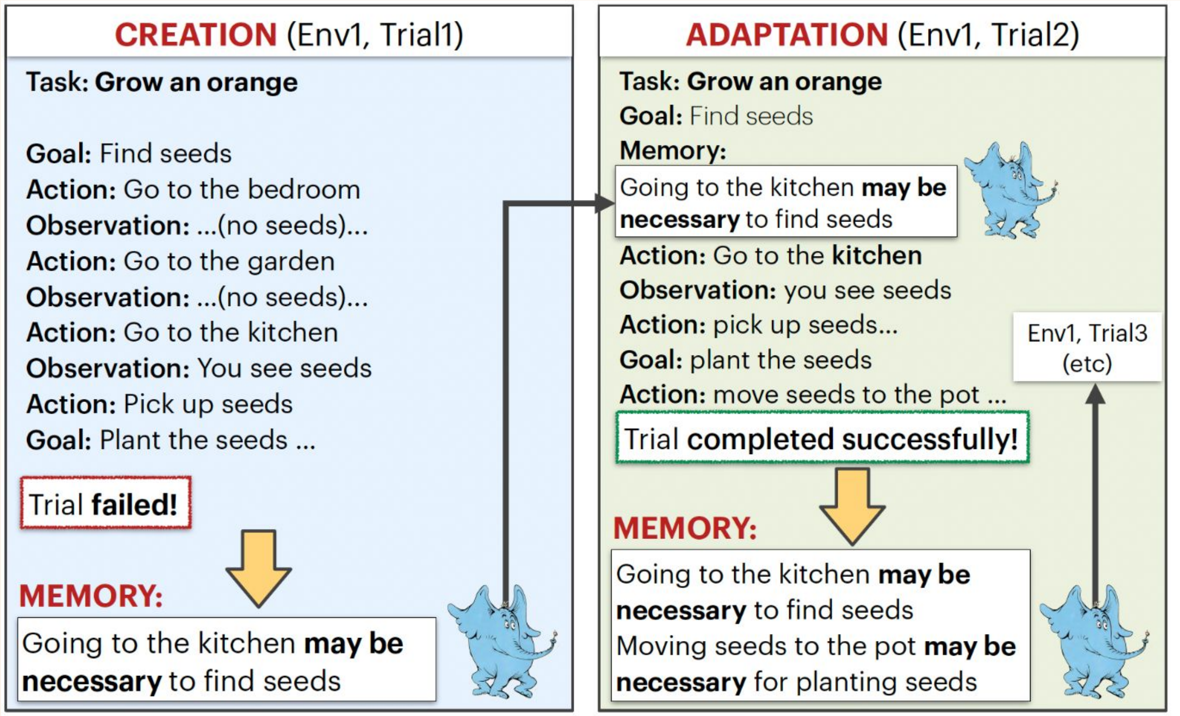


CLIN beats reflective SOTA

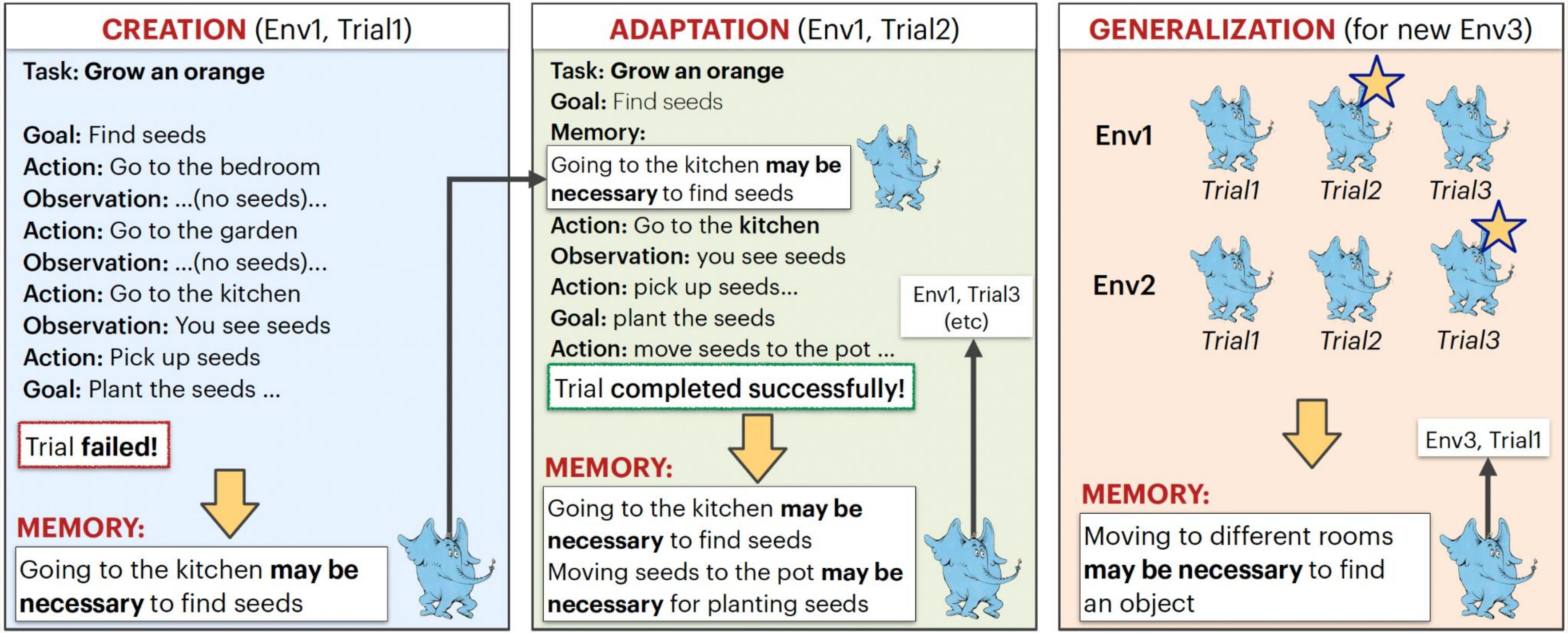


Ai2

Are Learned Hypotheses Generalizable?



Are Learned Hypotheses Generalizable?



Meta-Hypotheses

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

Meta-Hypotheses

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

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

Meta-Hypotheses

Task- and environment-specific memory cannot help generalize such as knowing how to boil water may not help knowing how to boil cadmium 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”

Ai2

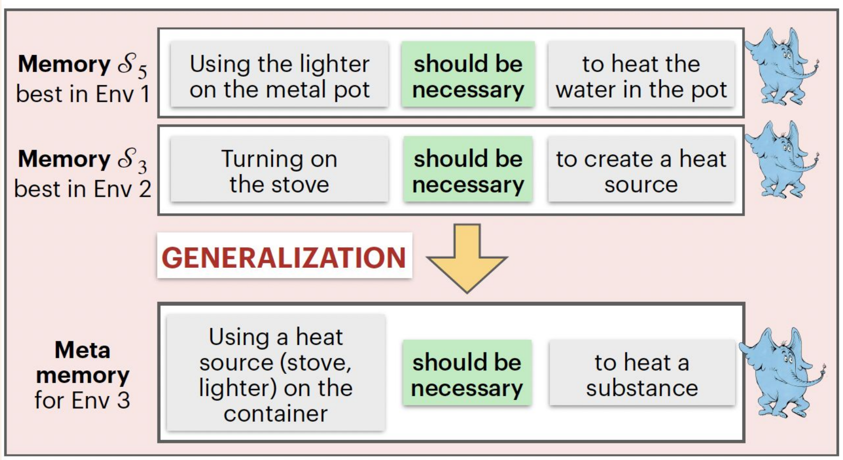
Meta-Hypotheses

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 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**

CLIN Generalizes to Novel Tasks

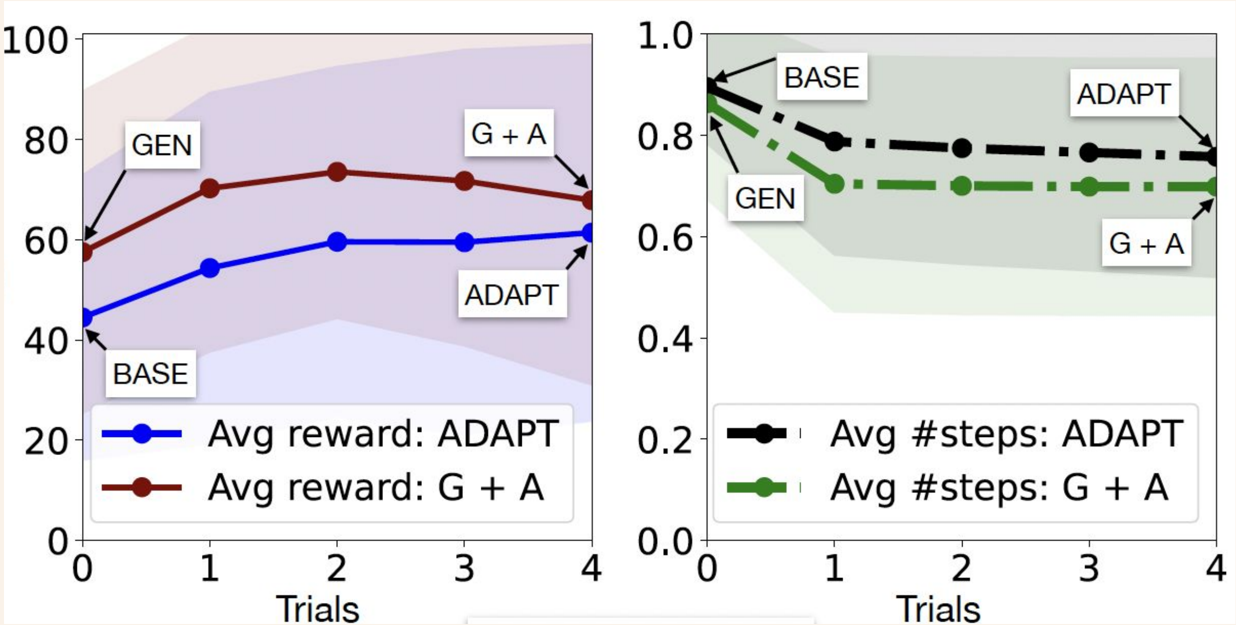
Train (in Env 1):
Boil water
Boil apple juice

Test (in Env 1):
Freeze Water

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

Performance gain in 38% episodes

Lesser steps



Precision of Learned Hypotheses

Natural selection of good hypotheses over time shows CLIN can auto-correct when the initial hypotheses are 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?

Hypothesis 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

Ai2

The Good and The Bad

CLIN is able to **compose hypotheses**

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

The Good and The Bad

CLIN is able to **compose hypotheses**

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**

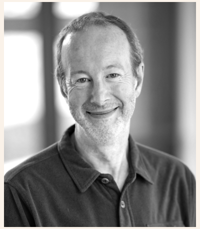
   If it has never visited an art studio,
   it will never "explore" to reach art studio for collecting paints

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

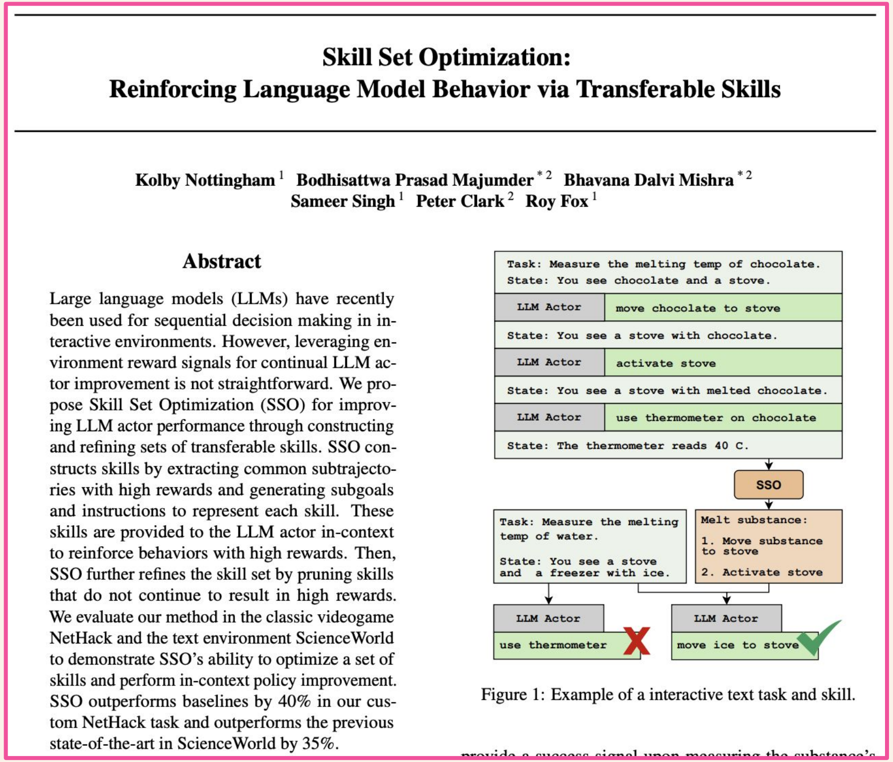







Hypothesis as a **skill**?

ICML, 2024



Skill Set Optimization: Reinforcing Language Model Behavior via Transferable Skills

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

Abstract

Large language models (LLMs) have recently been used for sequential decision making in interactive environments. However, leveraging environment reward signals for continual LLM actor improvement is not straightforward. We propose Skill Set Optimization (SSO) for improving LLM actor performance through constructing and refining sets of transferable skills. SSO constructs skills by extracting common subtrajectories with high rewards and generating subgoals and instructions to represent each skill. These skills are provided to the LLM actor in-context to reinforce behaviors with high rewards. Then, SSO further refines the skill set by pruning skills that do not continue to result in high rewards. We evaluate our method in the classic videogame NetHack and the text environment ScienceWorld to demonstrate SSO's ability to optimize a set of skills and perform in-context policy improvement. SSO outperforms baselines by 40% in our custom NetHack task and outperforms the previous state-of-the-art in ScienceWorld by 35%.

Figure 1: Example of a interactive text task and skill.

Skills

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

Ai2

Skill Definition

Target:
- goal state feature

Prerequisites:
- initial state features
- used for retrieval

Instructions:
- generic actions to execute

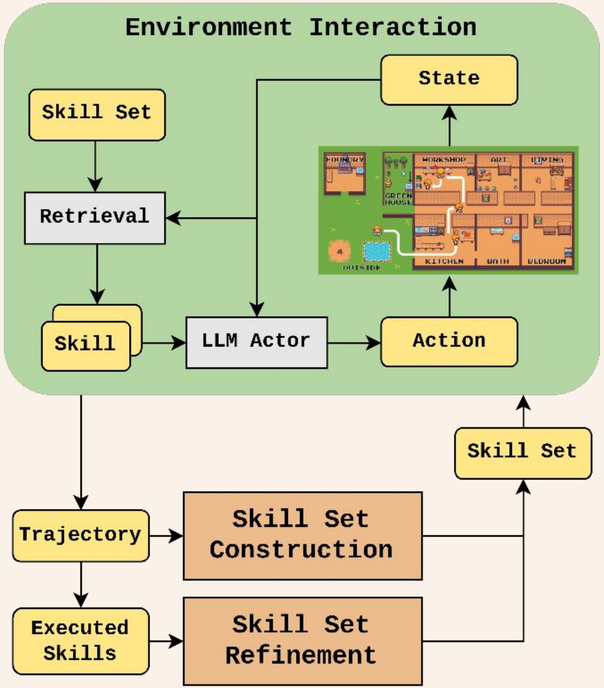
Example generated Skill

Target: agent is in the 'target location'

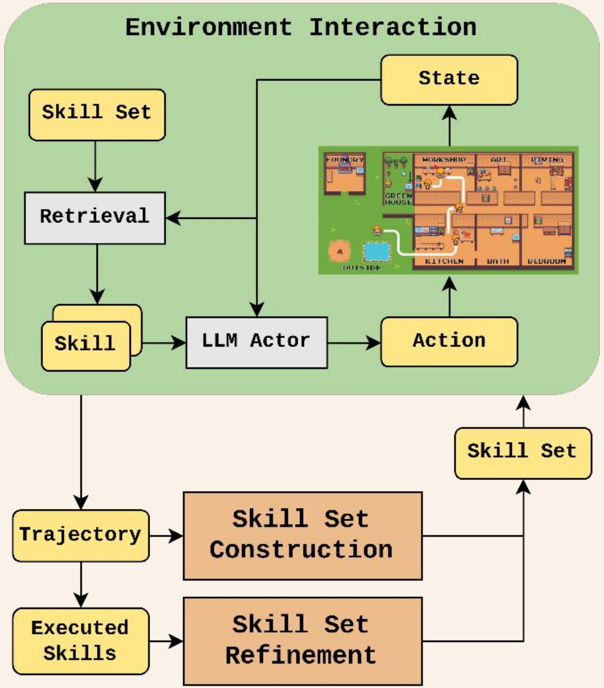
Prereqs:
1. agent is in a location that has a door leading to a hallway
2. there exists a known target location to which agent needs to move
3. agent is able to move (not restricted or blocked)

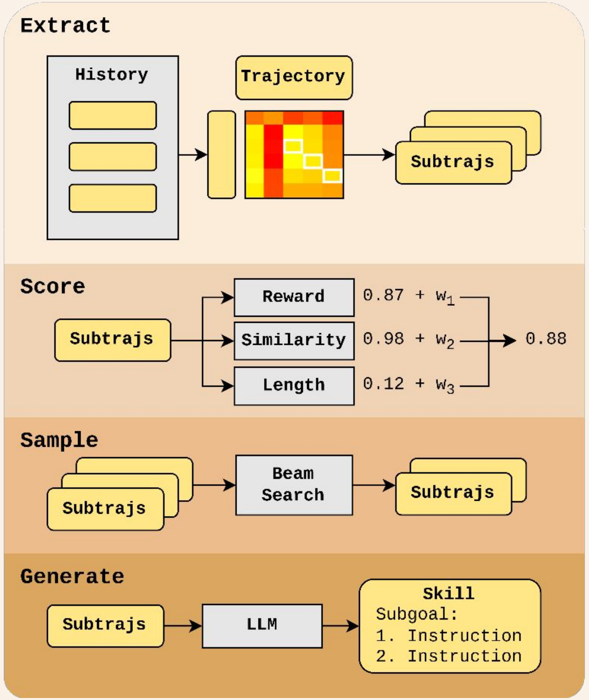
Instructions:
1. go to hallway
2. go to 'target location'

Skill Set Optimization



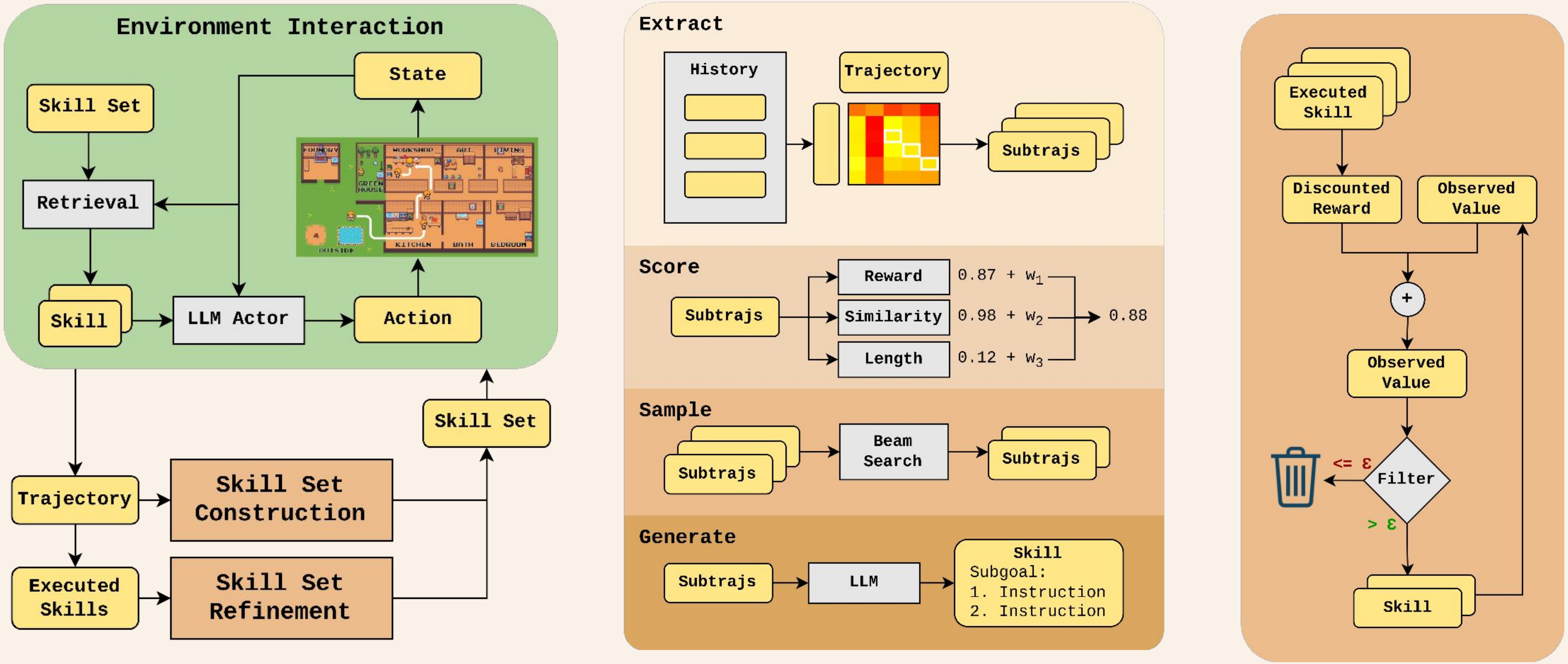
Skill Set Optimization





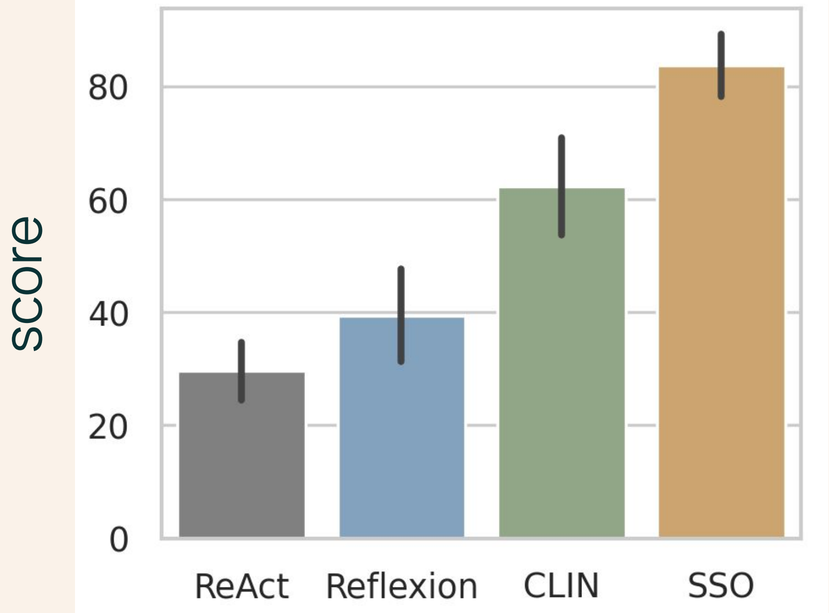
Ai2

Skill Set Optimization



Ai2

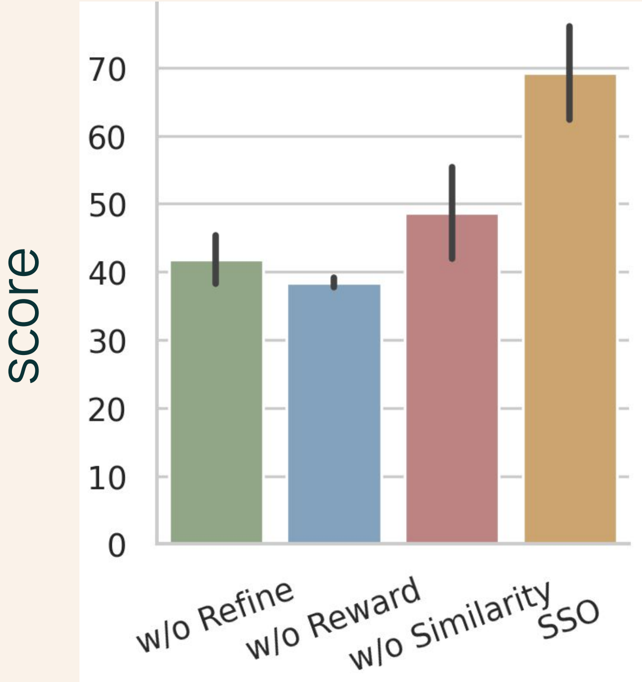
SSO Outperforms CLIN



ScienceWorld Task	ReAct	Adaptation Reflexion	Adaptation CLIN	Adaptation SSO	Transfer CLIN	Transfer 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**

Importance of Online Refinement

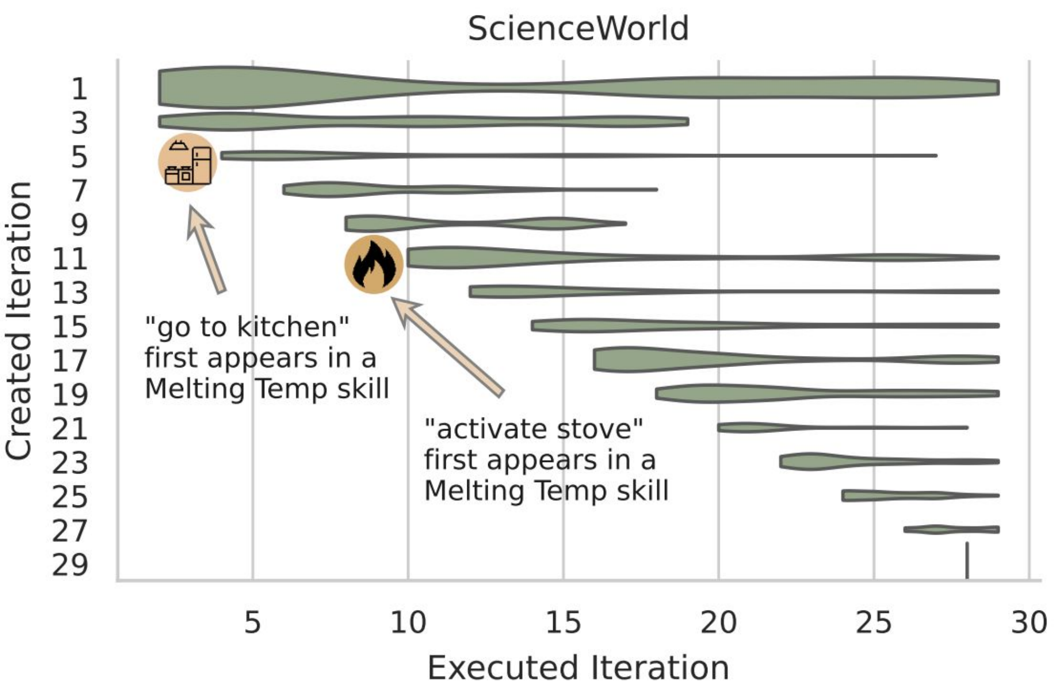
Importance of Online Refinement



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

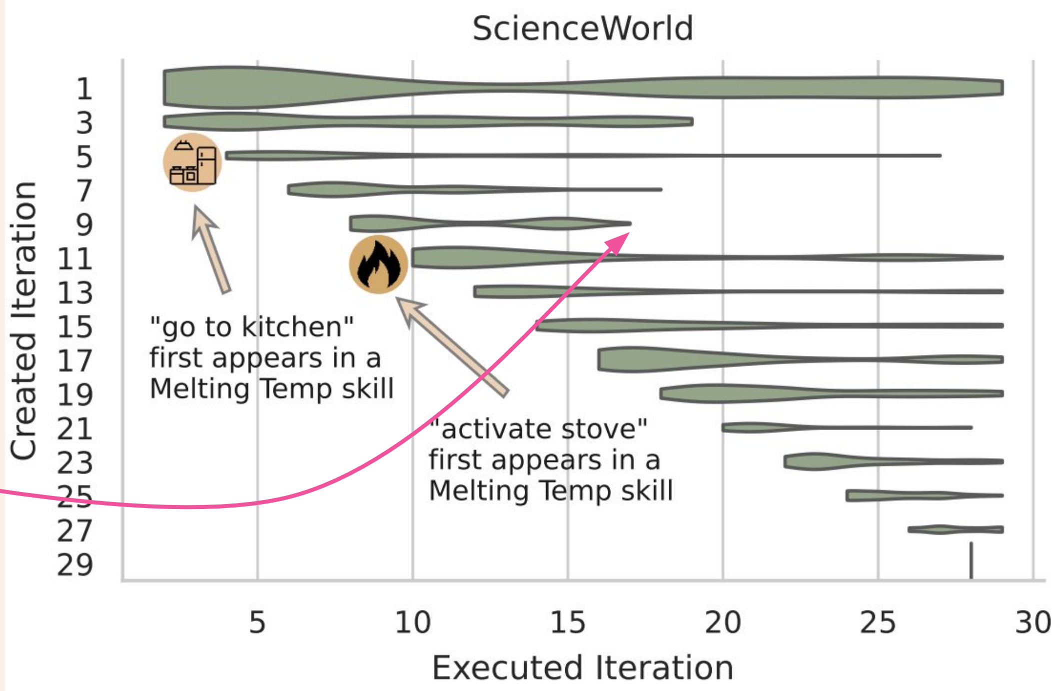
1. Skills are learned in order



Skill Lifecycle

1. Skills are learned in order

2. Old skills are forgotten in favor of the newer



The Process of Discovery

Communicative Intent

Recognizing the specific needs, intentions, and expectations of the task

Act in the World

Perform valid actions in the world to satisfy the communicative intent

Observe from the World

Observe the effect of performed action and validate with intent

Continual discovery of knowledge about the underlying world

E.g., Causal Facts, Sequential Effects, **Data-generating Fns**

Discovery, by nature, is sequential.

The process of discovery:

1. Conduct experiments to test pre-defined hypotheses

2. Observe & collect data; build methods to explain it

Methods of Scientific Inquiry

Theoretical Science

Develop models or theories to explain phenomena

Experimental Science

Conduct experiments to test pre-defined hypotheses

Observational Science

Observe & collect data, build methods to explain it

Methods of Scientific Inquiry

Theoretical Science

Develop models or theories to explain phenomena

Experimental Science

Conduct experiments to test pre-defined hypotheses

Observational Science

Observe & collect data, build methods to explain it

A lot of important science has come out of looking at **observational data.**

National Longitudinal Survey of Youth | 1979

NLSY79

U.S. BUREAU OF LABOR STATISTICS

500,000 results in S2 from 1979

Nurses' Health Study

BWH

37000+ papers published from 1976

Ai2

Methods of Scientific Inquiry

Theoretical Science

Develop models or theories to explain phenomena

Experimental Science

Conduct experiments to test pre-defined hypotheses

Observational Science

Observe & collect data, build methods to explain it

National Longitudinal Survey of Youth | 1979

NLSY79

U.S. BUREAU OF LABOR STATISTICS

500,000 results in S2 from 1979

A lot of important science has come out of looking at **observational data.**

Can we **autonomously** discover

- insights from datasets to reduce turnaround time?
- undiscovered knowledge without performing additional data collection?

Nurses' Health Study

BWH

37000+ papers published from 1976

Data-driven Discovery

- Comprehensive data-understanding
- Ex-ante hypothesis search/generation
- Planning & orchestrating research pathways
- Execute & verify candidate hypotheses
- Accommodating human feedback
- Reproducible and robust results

Data-driven Discovery

- Comprehensive data-understanding
- Ex-ante hypothesis search/generation
- Planning & orchestrating research pathways
- Execute & verify candidate hypotheses
- Accommodating human feedback
- Reproducible and robust results

Data-driven Discovery: Following Newell & Simon (1976), we define a **heuristic search problem** that aims to describe a given set of observations by uncovering the laws that govern its data-generating process.

E.g., "under context c, variables v have relationship r"

Newell, A. and Simon, H. A. Computer science as empirical inquiry: symbols and search. Commun. ACM, 1976

Automated Discovery in Past

	MLAgentBench	CoScientist	Bacon	DataLume	ThoughtSpot	Google AutoML	WolframAlpha*
Objective	Build ML optimal models autonomously	Autonomously plan, execute chemistry experiments	A Production system that discovers empirical laws	Explore data analyst support AI systems can provide	Data plotting, exploration with natural language	Builds optimal black-box model to serve at scale	Automatically analyze data
Comprehensive Data Understanding	Limited to model building	Targeted to Chemical Synthesis	N/A	Data understanding, Transformation	Data Scale only	Data Scale only	Limited to fixed datasets
Hypothesis Generation	N/A	Connect Data and Chemistry Papers, N/A on initial hypothesis	Heuristic-search on data leading to laws or equations	Partially with initial hypothesis	Partially with visualization	N/A	Partially with data analysis
Planning and Orchestrating Research Pathways	Yes, for model performance improvement	Plans for chemical synthesis	N/A, mostly heuristic driven	High-level planning w/o actionable steps	No	No	No
Hypothesis Evaluation	Verification with model performance and LLM efficiency	Conducts physical experiments	Basic heuristic calculations	Verification by interpreting statistical models	Partially with data exploration, visualization	Partially with feature importance	Partially with data analysis
Measurement of Progress	Intrinsic evaluation, but not with human feedback	Accommodates human feedback	N/A	N/A	N/A	Intrinsic model evaluation after training	No
Knowledge Integration	No	Knowledge from web and documents	N/A	Knowledge from LLMs	N/A	N/A	No

Ai2

ICML, 2024

Data-driven Discovery with Large Generative Models

Bodhisattwa Prasad Majumder [*1] **Harshit Surana** [*2] **Dhruv Agarwal** [3] **Sanchaita Hazra** [4]
Ashish Sabharwal [1] **Peter Clark** [1]

Abstract

With the accumulation of data at an unprecedented rate, its potential to fuel scientific discovery is growing exponentially. This position paper urges the Machine Learning (ML) community to exploit the capabilities of large generative models (LGMs) to develop automated systems for end-to-end *data-driven discovery*—a paradigm encom-

sions (Bianchini et al., 2022). To facilitate future scientific progress, it is, therefore, imperative to develop automated systems that are capable of continuous ingestion, creative generation, and analytical reasoning at a massive scale.

Developing an end-to-end discovery system is challenging. Previous works have either severely lacked the requisite computational power (Langley, 1981; Langley et al., 1984; 1983), developed domain-specific bespoke methodologies

DiscoveryBench: Towards Data-Driven Discovery with Large Language Models

Bodhisattwa Prasad Majumder[*1] **Harshit Surana**[*12] **Dhruv Agarwal**[*3]
Bhavana Dalvi Mishra[*1] **Abhijeetsingh Meena**[2] **Aryan Prakhar**[2] **Tirth Vora**[2]
Tushar Khot[1] **Ashish Sabharwal**[1] **Peter Clark**[1]
[1]Allen Institute for AI [2]OpenLocus [3]University of Massachusetts Amherst

Website: `https://github.com/allenai/discoverybench`
🤗 `https://huggingface.co/datasets/allenai/discoverybench`
[*]equal contributions

Abstract

Can the rapid advances in code generation, function calling, and data analysis using large language models (LLMs) help automate the search and verification of hypotheses purely from a set of provided datasets? To evaluate this question, we present DiscoveryBench, the first comprehensive benchmark that formalizes the multi-step process of data-driven discovery. The benchmark is designed to systematically assess current model capabilities in discovery tasks and provide a useful resource for improving them. Our benchmark contains 264 tasks collected across 6 diverse domains, such as sociology and engineering, by manually deriving discovery workflows from published papers to approximate the real-world challenges faced by researchers, where each task is defined by a dataset, its metadata, and a discovery goal in natural language. We additionally provide 903 synthetic tasks to conduct controlled evaluations across task complexity. Furthermore, our structured formalism of data-driven discovery enables a facet-based evaluation that provides useful insights into different failure modes. We evaluate several popular LLM-based reasoning frameworks using both open and closed LLMs as baselines on DiscoveryBench and find that even the best system scores only 25%. Our benchmark, thus, illustrates the challenges in autonomous data-driven discovery and serves as a valuable resource for the community to make progress.

Ai2

Data-driven Discovery as a Predictive Task

Given a dataset **D** and a Discovery Goal **G**, derive the most specific hypothesis **H** addressing G and supported by D.

Alternatively,
A data-driven hypothesis **H** is a declarative sentence about the state of the world whose truth value may be inferred from a given dataset D using a verification procedure V: H → {supported, unsupported}, for instance, via *statistical modeling*.

Data-driven Discovery as a Predictive Task

Given a dataset **D** and a Discovery Goal **G**, derive the most specific hypothesis **H** addressing G and supported by D.

Alternatively,
A data-driven hypothesis **H** is a declarative sentence about the state of the world whose truth value may be inferred from a given dataset D using a verification procedure V: H → {supported, unsupported}, for instance, via *statistical modeling*.

Inspired by Thompson and Skau (2023), we introduce a structured formalism that breaks a hypothesis down into **three hypothesis dimensions**:

Context: **Boundary conditions** that limit the scope of a hypothesis. E.g., “for men over the age of 30”

Variables: **Known set of concepts** that interact in a meaningful way under a given context to produce the hypothesis. E.g., gender, age, or income

Relationship: **Interactions between a given set of variables** under a given context that produces the hypothesis. E.g., “quadratic relationship”, “inversely proportional”, or piecewise conditionals

W. H. Thompson and S. Skau. On the scope of scientific hypotheses. Royal Society Open Science, 2023

Data-driven Discovery as a Predictive Task

Given a dataset **D** and a Discovery Goal **G**, derive the most specific hypothesis **H** addressing G and supported by D.

Alternatively,
A data-driven hypothesis **H** is a declarative sentence about the state of the world whose truth value may be inferred from a given dataset D using a verification procedure V: H → {supported, unsupported}, for instance, via *statistical modeling*.

Inspired by Thompson and Skau (2023), we introduce a structured formalism that breaks a hypothesis down into **three hypothesis dimensions**:

Context: Boundary conditions that limit the scope of a hypothesis. E.g., "for men over the age of 30"

Variables: Known set of concepts that interact in a meaningful way under a given context to produce the hypothesis. E.g., gender, age, or income

Relationship: Interactions between a given set of variables under a given context that produces the hypothesis. E.g., "quadratic relationship", "inversely proportional", or piecewise conditionals

W. H. Thompson and S. Skau. On the scope of scientific hypotheses. Royal Society Open Science, 2023

Dataset:

habitat type	nonnative gardening	nonnative unintentional	nonnative agriforest	elevation	...
croplands	5	0	2	675	
wetlands	0	4	1	88	
urban	2	1	0	329	
...	...	...	...	...	

Goal: How did urban land use affect the invasion of different types of introduced plants in Catalonia?

	gold	predicted	score
context	urban habitat type	urban habitat type	1.0
variable	gardening, unintentional	gardening, agriforst	0.3
relationship	reduced	increased	0.0

Final Score: 0.21

Urban land use reduced invasion by gardening plants over unintentionally introduced ones.

Ai2

DiscoveryBench

264 Tasks, 20+ papers, 6 domains

We replicate the **scientific process** undertaken by researchers to search for and validate a hypothesis from datasets

DiscoveryBench

264 Tasks, 20+ papers, 6 domains

We replicate the **scientific process** undertaken by researchers to search for and validate a hypothesis from datasets

Data-first: Filter papers + workflows based on public datasets: National Longitudinal Surveys, Global Biodiversity Info Facility, World Bank Open Data; 2) replicate in Python.

Replication took up to 90 person-hours per dataset, often (30%) not resulting in success.

Code-first: Checked 785 repos + datasets, 85% had missing or non-adaptable code to Python, or closed datasets. Only few passed the check.

Papers from Nature, AER, etc.

DiscoveryBench

264 Tasks, 20+ papers, 6 domains

We replicate the **scientific process** undertaken by researchers to search for and validate a hypothesis from datasets

Data-first: Filter papers + workflows based on public datasets: National Longitudinal Surveys, Global Biodiversity Info Facility, World Bank Open Data; 2) replicate in Python.

Replication took up to 90 person-hours per dataset, often (30%) not resulting in success.

Code-first: Checked 785 repos + datasets, 85% had missing or non-adaptable code to Python, or closed datasets. Only few passed the check.

Papers from Nature, AER, etc.

Task Dataset: Dataset contains information from National Longitudinal Survey of Youth (NLSY79). It includes information about the Demographics, Family Background, Education ...

Discovery Goal: How does socioeconomic status affect the likelihood of completing a BA degree?

Target Hypothesis:
Socioeconomic status has a positive relationship with college degree completion with a coefficient of 0.4729 with statistical significance.

DiscoveryBench

264 Tasks, 20+ papers, 6 domains

We replicate the **scientific process** undertaken by researchers to search for and validate a hypothesis from datasets

Data-first: Filter papers + workflows based on public datasets: National Longitudinal Surveys, Global Biodiversity Info Facility, World Bank Open Data; 2) replicate in Python.

Replication took up to 90 person-hours per dataset, often (30%) not resulting in success.

Code-first: Checked 785 repos + datasets, 85% had missing or non-adaptable code to Python, or closed datasets. Only few passed the check.

Papers from Nature, AER, etc.

Task Dataset: Dataset contains information from National Longitudinal Survey of Youth (NLSY79). It includes information about the Demographics, Family Background, Education …

Discovery Goal: How does socioeconomic status affect the likelihood of completing a BA degree?

Target Hypothesis: Socioeconomic status has a positive relationship with college degree completion with a coefficient of 0.4729 with statistical significance.

Data Loading & Cleaning

```python
df=pd.read_csv('NLSCombine.csv')

columns_to_clean = ['Number of students in class last year attended at this school, 1981',"Highest grade completed by respondent's mother, 1979",
        "Highest grade completed by respondent's father, 1979",  'Family size, 1979',
        'Highest grade completed, 1979', 'Rank in class last year attended at this school, 1981',
         'ABILITY', 'Total net family income, previous calendar year, 1979']

for col in columns_to_clean:
    df[col] = df[col].apply(lambda x: np.nan if x < 0 else x)

df = df.dropna(subset=columns_to_clean, thresh=6)

# Standardize the continuous and ordinal variables
scaler = StandardScaler()
df[['STANDARDIZED_INCOME', 'STANDARDIZED_FAMILY_SIZE',
'STANDARDIZED_FATHER_EDUCATION', 'STANDARDIZED_MOTHER_EDUCATION']] = scaler.
fit_transform(
    df[['Total net family income, previous calendar year, 1979','Family size, 1979' ,
    "Highest grade completed by respondent's father, 1979", "Highest grade completed
    by respondent's mother, 1979"]])

# Initialize the IterativeImputer
imputer = IterativeImputer(max_iter=10, random_state=0)

# Columns to impute - focusing on the ones with NaN values and relevant to SES
calculation
```

Calculate Academic Ability et al.

```python
score_cols=[
    'ASVAB - Arithmetic Reasoning Z Score (rounded), 1981',
    'ASVAB - Word Knowledge Z Score (rounded), 1981',
    'ASVAB - Paragraph Comprehension Z Score (rounded), 1981',
    'ASVAB - Mathematics Knowledge Z Score (rounded), 1981'
    ]

df['all scores exist']=df[score_cols].gt(0).all(axis=1)

df['ABILITY'] =
    df['ASVAB - Arithmetic Reasoning Z Score (rounded), 1981'] +
    df['ASVAB - Word Knowledge Z Score (rounded), 1981'] +
    df['ASVAB - Paragraph Comprehension Z Score (rounded), 1981'] +
    df['ASVAB - Mathematics Knowledge Z Score (rounded), 1981']

df['BA DEGREE COMPLETED'] = df['Highest grade completed, 1979'].gt(14)
```

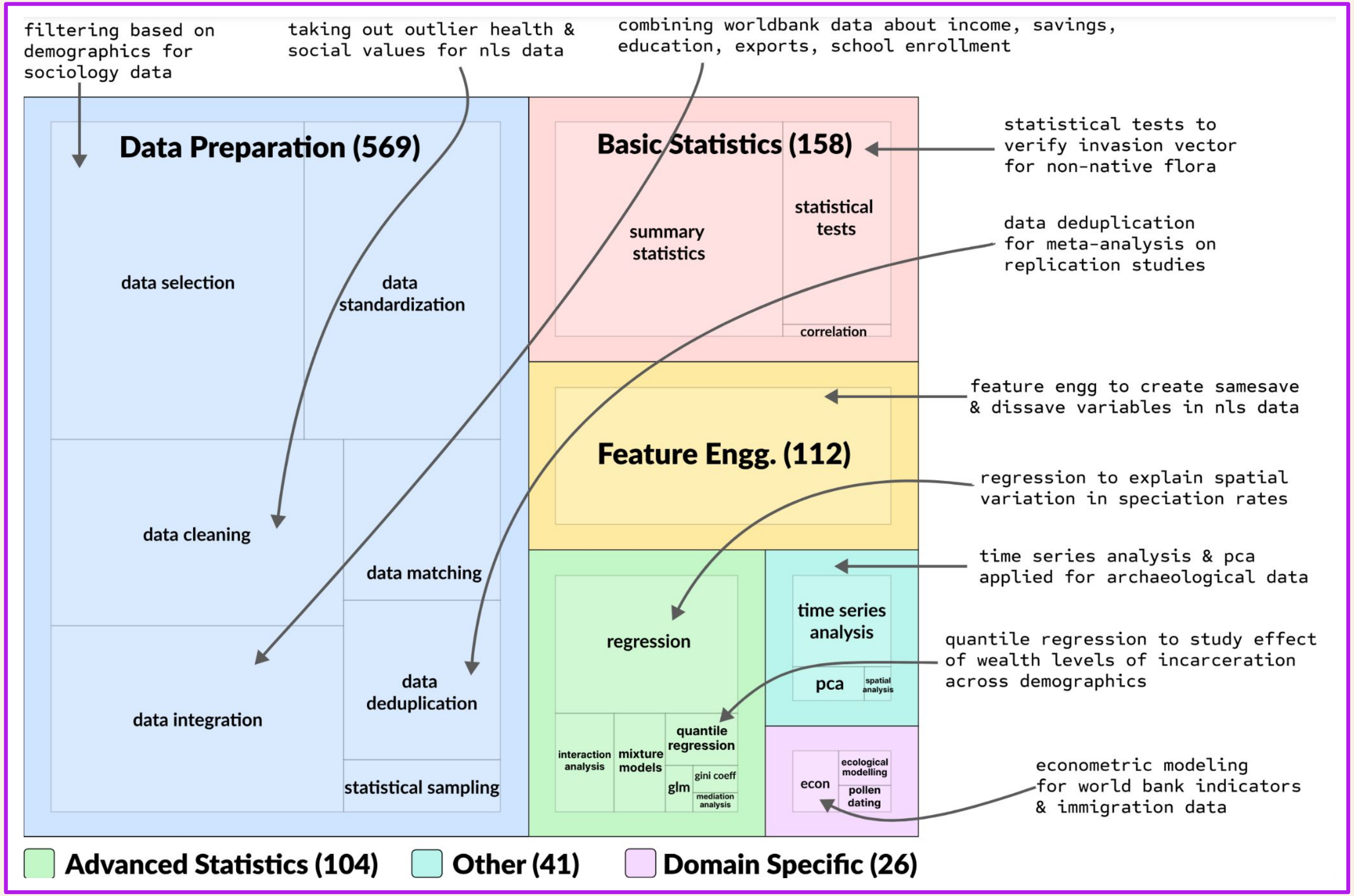
Calculate SES

```python
weight_income = 0.5
weight_education = 0.25

# Calculate SES as a weighted composite of standardized measures
df['SES'] = (
    weight_income * df['STANDARDIZED_INCOME'] +
    weight_education * df['STANDARDIZED_FATHER_EDUCATION'] +
    weight_education * df['STANDARDIZED_MOTHER_EDUCATION']
)
```

Data Filtering & Modeling

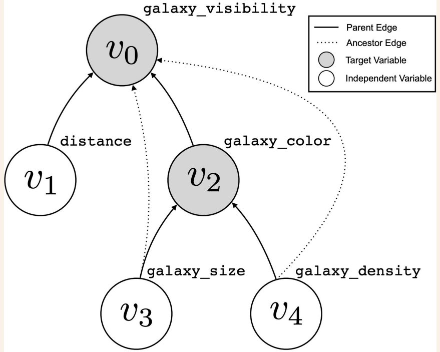
```python
# Limit data to records with percentile in the range 0 to 100
df = df[(df['PERCENTILE IN CLASS'] <= 100) &
    (df['PERCENTILE IN CLASS'] >= 0)]

sub_dataset['BA DEGREE COMPLETED'] =
    sub_dataset['BA DEGREE COMPLETED'].astype(int)

X = sub_dataset[['SES']]
X = sm.add_constant(X)
y = sub_dataset['BA DEGREE COMPLETED']

model = sm.Logit(y, X)
result = model.fit()

result.summary()
```

DB-Real (6 domains: sociology, biology, humanities, economics, engineering, & meta-science)



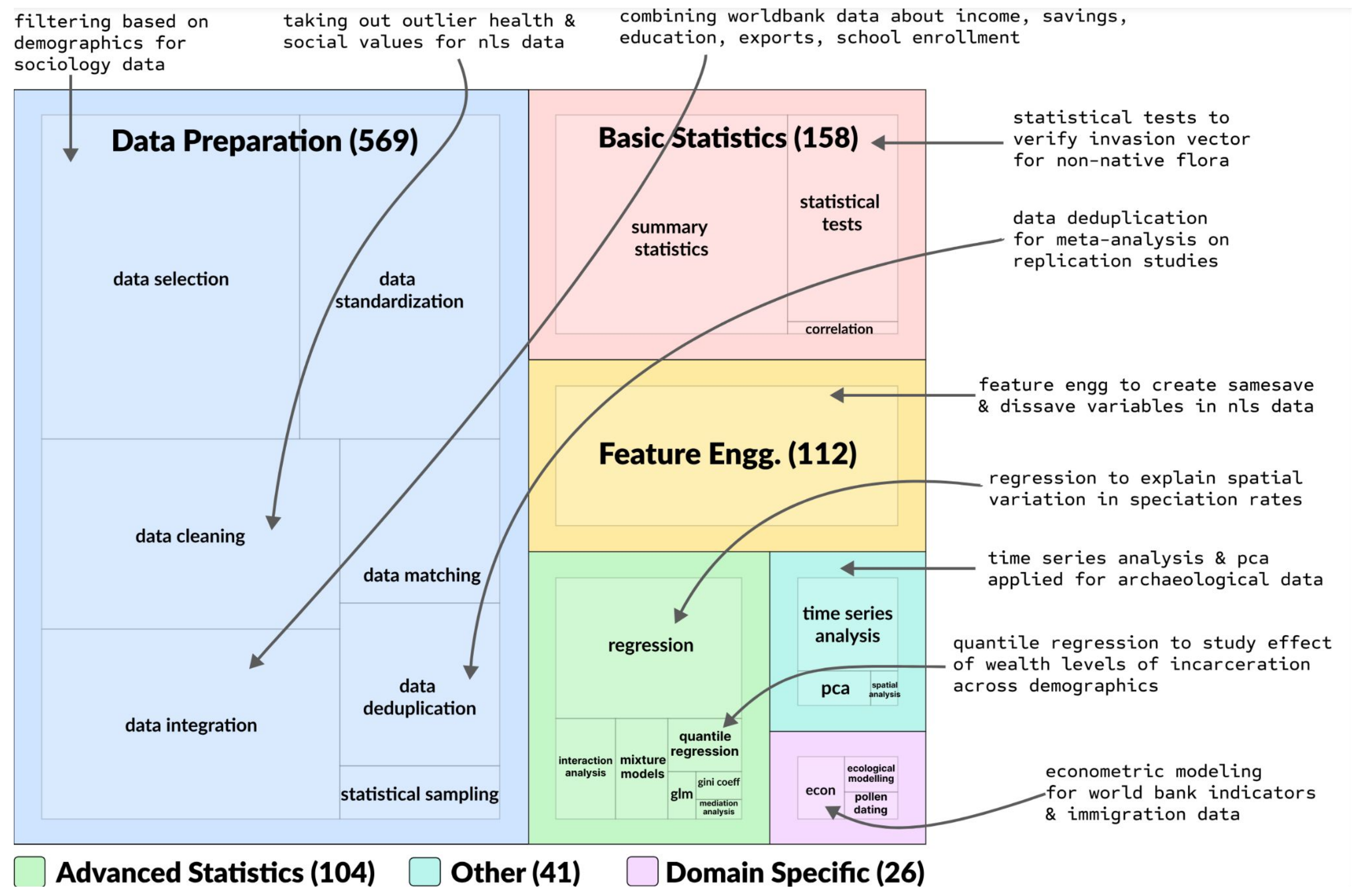
Ai2

Hypothesis Semantic Tree

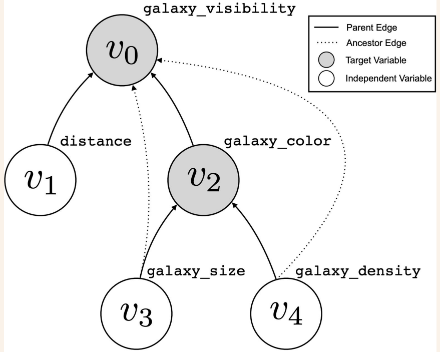


Task difficulty $\propto$ max path length between obs. and target node

DB-Real (6 domains: sociology, biology, humanities, economics, engineering, & meta-science)



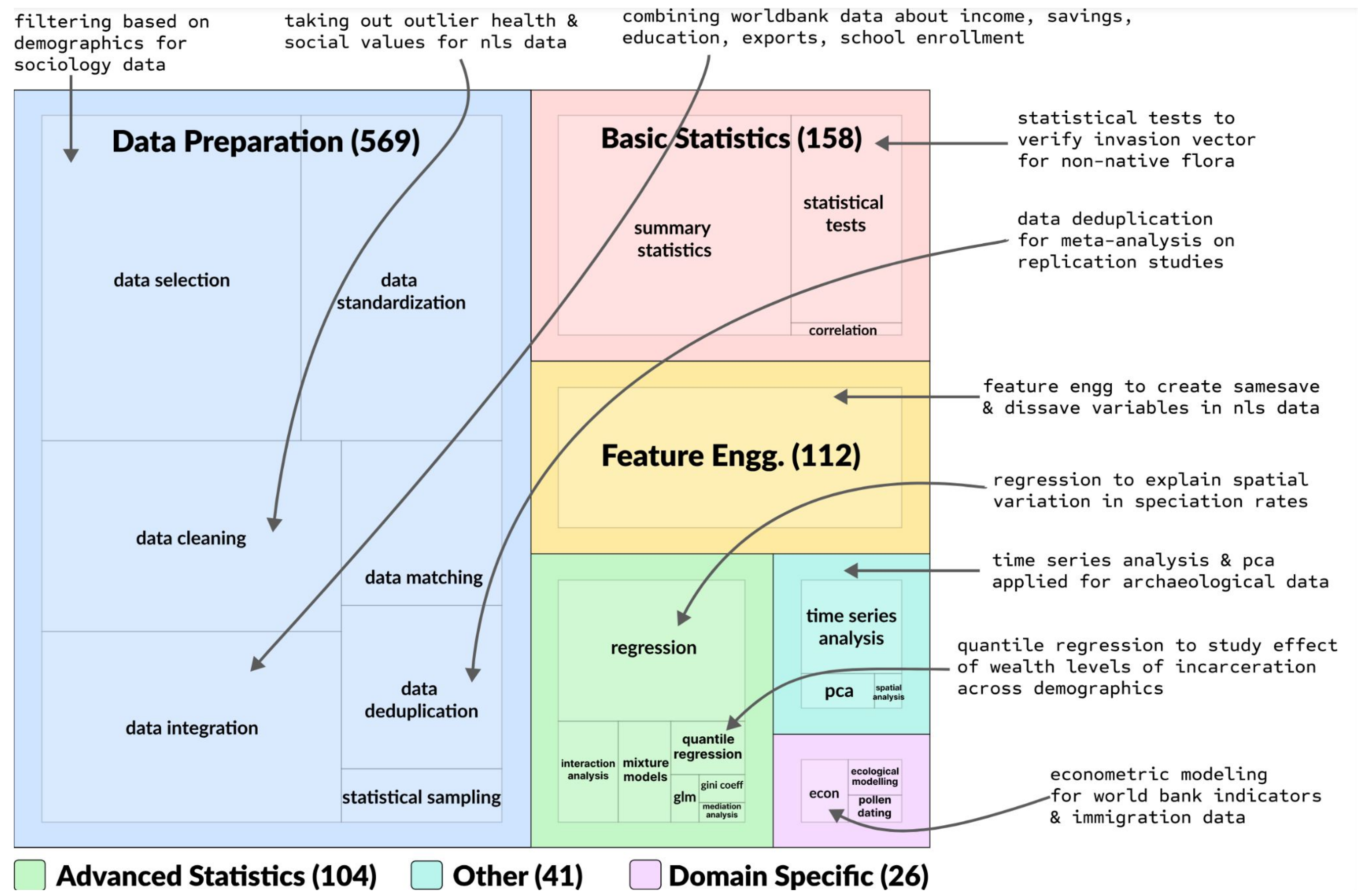
Hypothesis Semantic Tree



Task difficulty ∝ max path length between obs. and target node

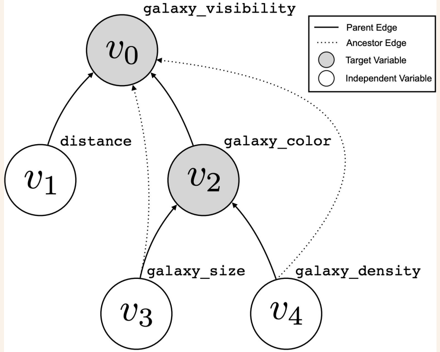
	Train	Test
# tasks	25	239
# unique hypotheses	14	144
# tasks need > 1 dataset	4	110
# domains	3	6

DB-Real (6 domains: sociology, biology, humanities, economics, engineering, & meta-science)



Ai2

Hypothesis Semantic Tree



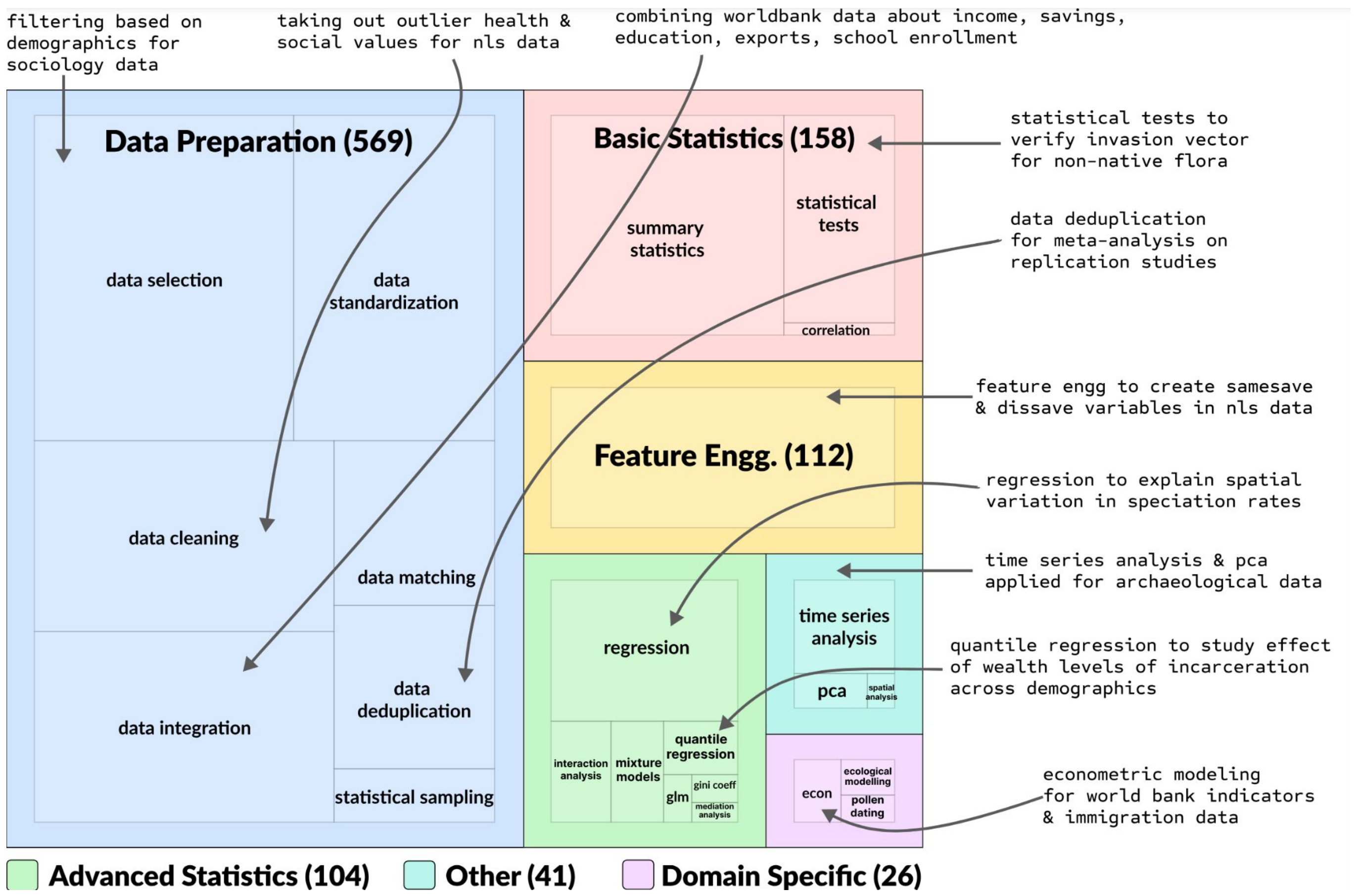
Task difficulty $\propto$ max path length between obs. and target node

	Train	Test
# tasks	25	239
# unique hypotheses	14	144
# tasks need > 1 dataset	4	110
# domains	3	6

DB-Synth *(skipped)*

LLM-based semantic tree construction + data generation

DB-Real (6 domains: sociology, biology, humanities, economics, engineering, & meta-science)



Discovery Agents

All discovery agents have access to a python environment, capable of generating and executing programs on the datasets

CodeGen

generates the entire code at one go to solve the task, with help of a demonstration example in the context.

After code execution and based on the result, it generates the NL hypothesis and summarizes the workflow

ReAct

solves the task by generating thought and subsequent codes in a multi-turn fashion.

A traditional sequential-decision maker.

DataVoyager

is a multi-component data-driven discovery agent.

It has four components: planner, code generator, data analysis, and critic, that orchestrate the discovery process.

Reflexion (Oracle)

is an extension of CodeGen agent, where at the end of one trial, we provide an “oracle” feedback about task completion, and it generates a reflection to improve in the next trial till it solves the task, or maximum trials (3) are reached.

Can Discovery Agents Solve Discovery Tasks?

All discovery agents have access to a python environment, capable of generating and executing programs on the datasets

	GPT-4o	GPT-4p	Llama-3
DB-REAL			
NoDataGuess	0.0	4.7	11.5
CodeGen	15.5	16.3	12.1
React	15.4	15.6	13.5
DataVoyager	15.4	13.9	11.5
Reflexion (Oracle)	**24.5**	**19.5**	**22.5**

Overall performance for all framework-LLM pairs is low

Llama-3 is almost equally performant with GPTs

Can Discovery Agents Solve Discovery Tasks?

All discovery agents have access to a python environment, capable of generating and executing programs on the datasets

	GPT-4o	GPT-4p	Llama-3
DB-REAL			
NoDataGuess	0.0	4.7	11.5
CodeGen	15.5	16.3	12.1
React	15.4	15.6	13.5
DataVoyager	15.4	13.9	11.5
Reflexion (Oracle)	**24.5**	**19.5**	**22.5**

With oracle Reflexion, performance significantly improves.

Agents' performance could improve with human-in-the-loop

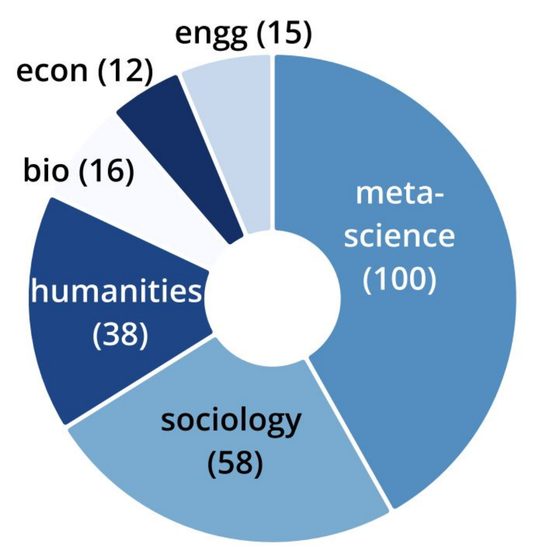
Can LLMs "Cheat" by Hallucination?

	GPT-4o	GPT-4p	Llama-3
DB-Real			
NoDataGuess	0.0	4.7	11.5
CodeGen	15.5	16.3	12.1
React	15.4	15.6	13.5
DataVoyager	15.4	13.9	11.5
Reflexion (Oracle)	**24.5**	**19.5**	**22.5**

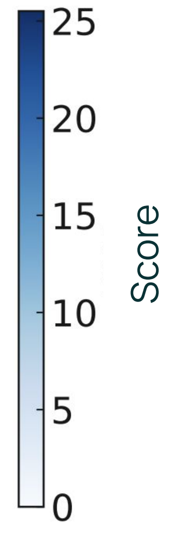
We set up agents to generate the final hypothesis only with the task and data description, but without provisioning any data!

Llama-3 performs similarly in both data and no-data modes!!

Graded Performance of the Best Agent

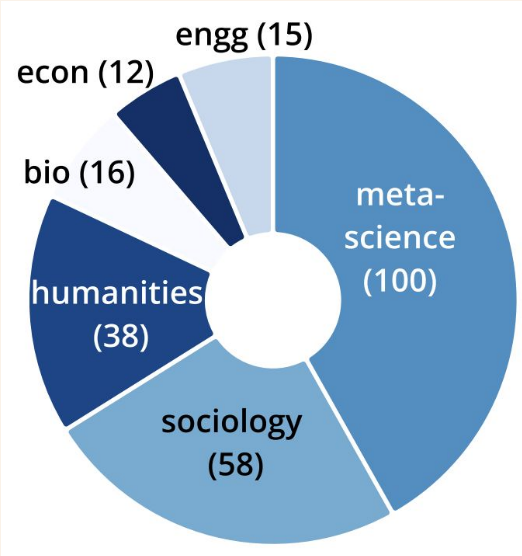


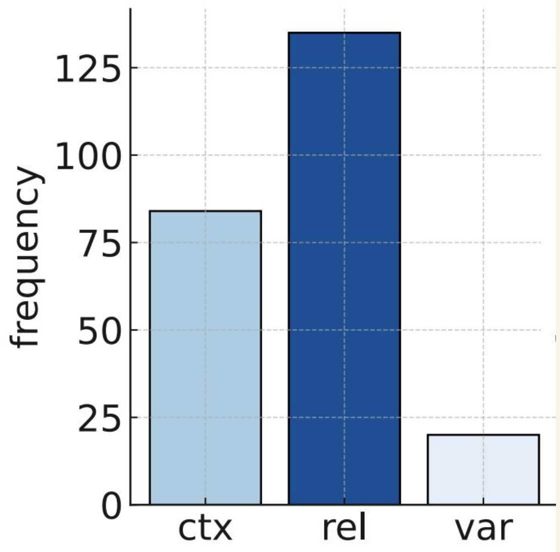
Biology (0%) and engineering (7%) perform the worst. They require advanced stat methods. Economics (25%) and sociology (23%) perform better.

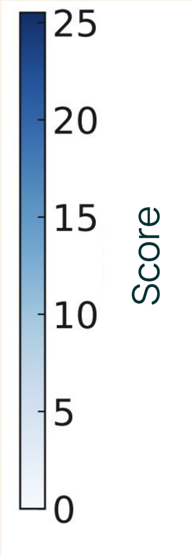


Ai2

Graded Performance of the Best Agent



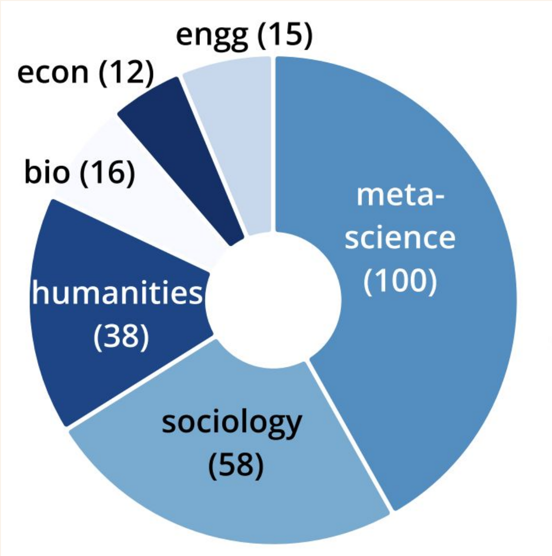


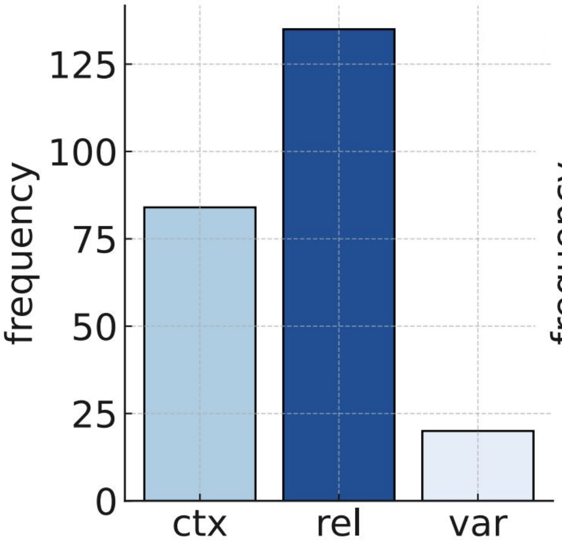


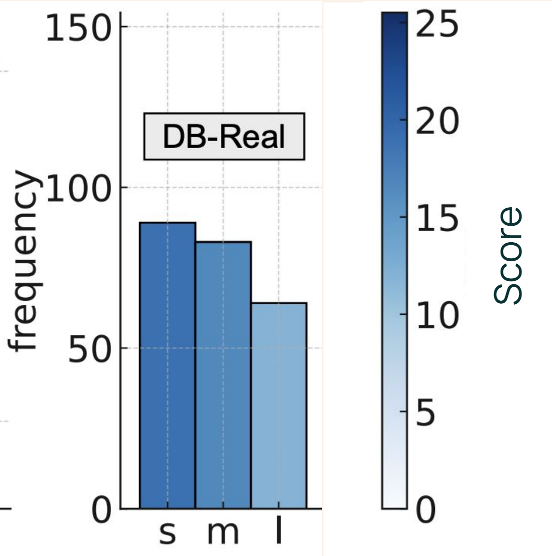
Biology (0%) and engineering (7%) perform the worst. They require advanced stat methods. **Economics (25%) and sociology (23%)** perform better.

Goals related to **discovering a relationship** given context and variables **are more easily solved** than the other two types of goals.

Graded Performance of the Best Agent







Biology (0%) and engineering (7%) perform the worst. They require advanced stat methods. **Economics (25%) and sociology (23%)** perform better.

Goals related to **discovering a relationship** given context and variables **are more easily solved** than the other two types of goals.

Decreasing trend in performance as workflow length increases. The performance drops significantly even for medium-length workflows.

Ai2

Summary

- Communicative agents form better hypotheses in structured formalism compared to generic abstractions
- Online refinement has net welfare benefit, imagine learning a library of hypothesis – automated knowledge base construction?
- It's possible to view automated discovery as a grounded interactive task and communicative agents can offer progress

@mbodhisattwa

@bodhisattwam@allenai.org

Thanks!

Ai2