



COS 484

Natural Language Processing

LI4: Language Agents

Spring 2026

Large Language Models are
everywhere

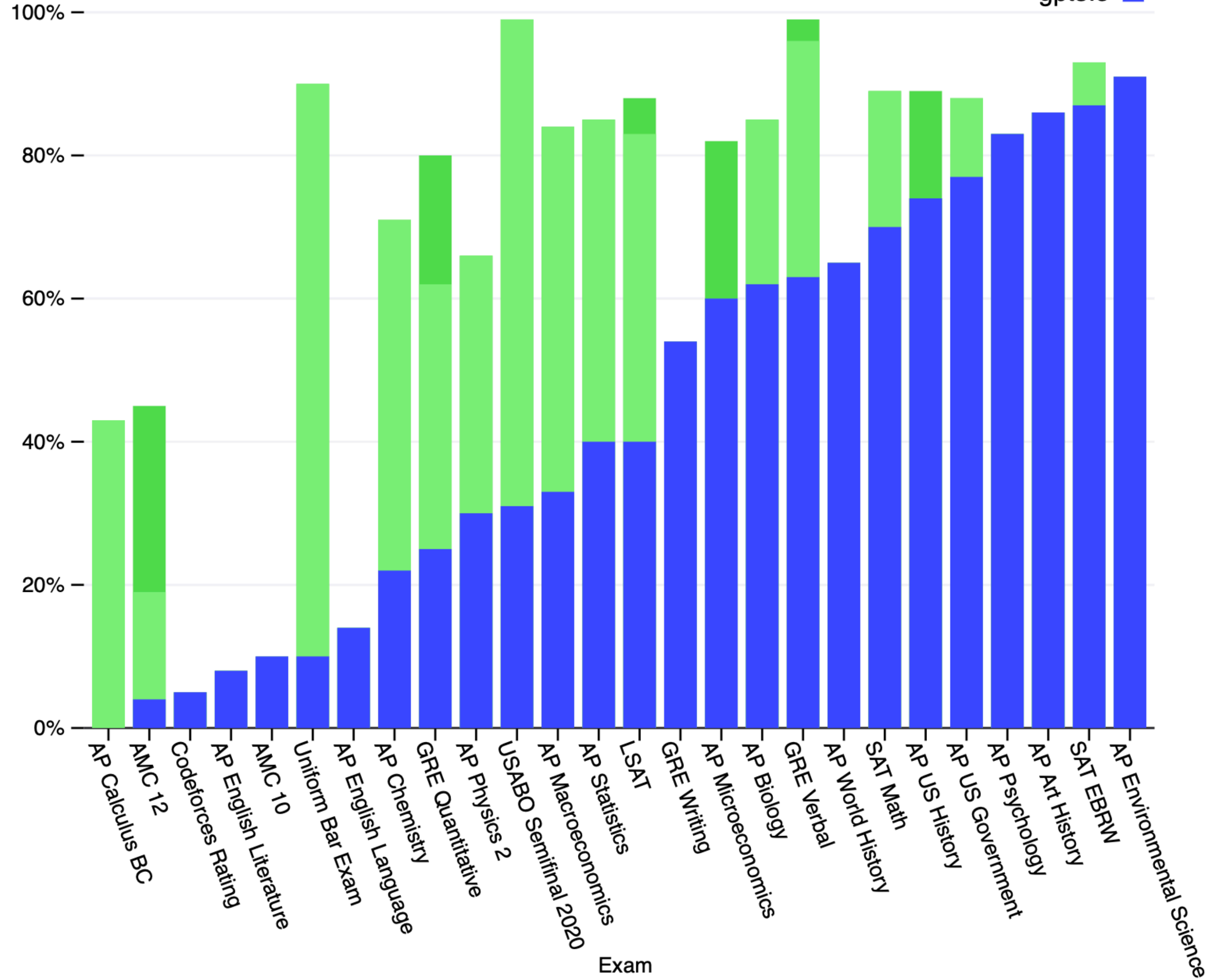
NLP Benchmarks

	GPT-4	GPT-3.5	LM SOTA	SOTA
	Evaluated few-shot	Evaluated few-shot	Best external LM evaluated few-shot	Best external model (incl. benchmark-specific tuning)
MMLU [49] Multiple-choice questions in 57 subjects (professional & academic)	86.4% 5-shot	70.0% 5-shot	70.7% 5-shot U-PaLM [50]	75.2% 5-shot Flan-PaLM [51]
HellaSwag [52] Commonsense reasoning around everyday events	95.3% 10-shot	85.5% 10-shot	84.2% LLaMA (validation set) [28]	85.6 ALUM [53]
AI2 Reasoning Challenge (ARC) [54] Grade-school multiple choice science questions. Challenge-set.	96.3% 25-shot	85.2% 25-shot	85.2% 8-shot PaLM [55]	86.5% ST-MOE [18]
WinoGrande [56] Commonsense reasoning around pronoun resolution	87.5% 5-shot	81.6% 5-shot	85.1% 5-shot PaLM [3]	85.1% 5-shot PaLM [3]
HumanEval [43] Python coding tasks	67.0% 0-shot	48.1% 0-shot	26.2% 0-shot PaLM [3]	65.8% CodeT + GPT-3.5 [57]
DROP [58] (F1 score) Reading comprehension & arithmetic.	80.9 3-shot	64.1 3-shot	70.8 1-shot PaLM [3]	88.4 QDGAT [59]
GSM-8K [60] Grade-school mathematics questions	92.0%* 5-shot chain-of-thought	57.1% 5-shot	58.8% 8-shot Minerva [61]	87.3% Chinchilla + SFT+ORM-RL, ORM reranking [62]

Exam results (ordered by GPT-3.5 performance)

Estimated percentile lower bound (among test

gpt-4
gpt-4 (no vision)
gpt3.5



LLMs can also enhance
autonomous *agents*

What is a language agent?

- Language model = probabilistic model of text
 - What does it provide?
- Language agent
 - An agent that can understand and generate language



 **a·gent**
/ˈɑj(ə)nt/

noun

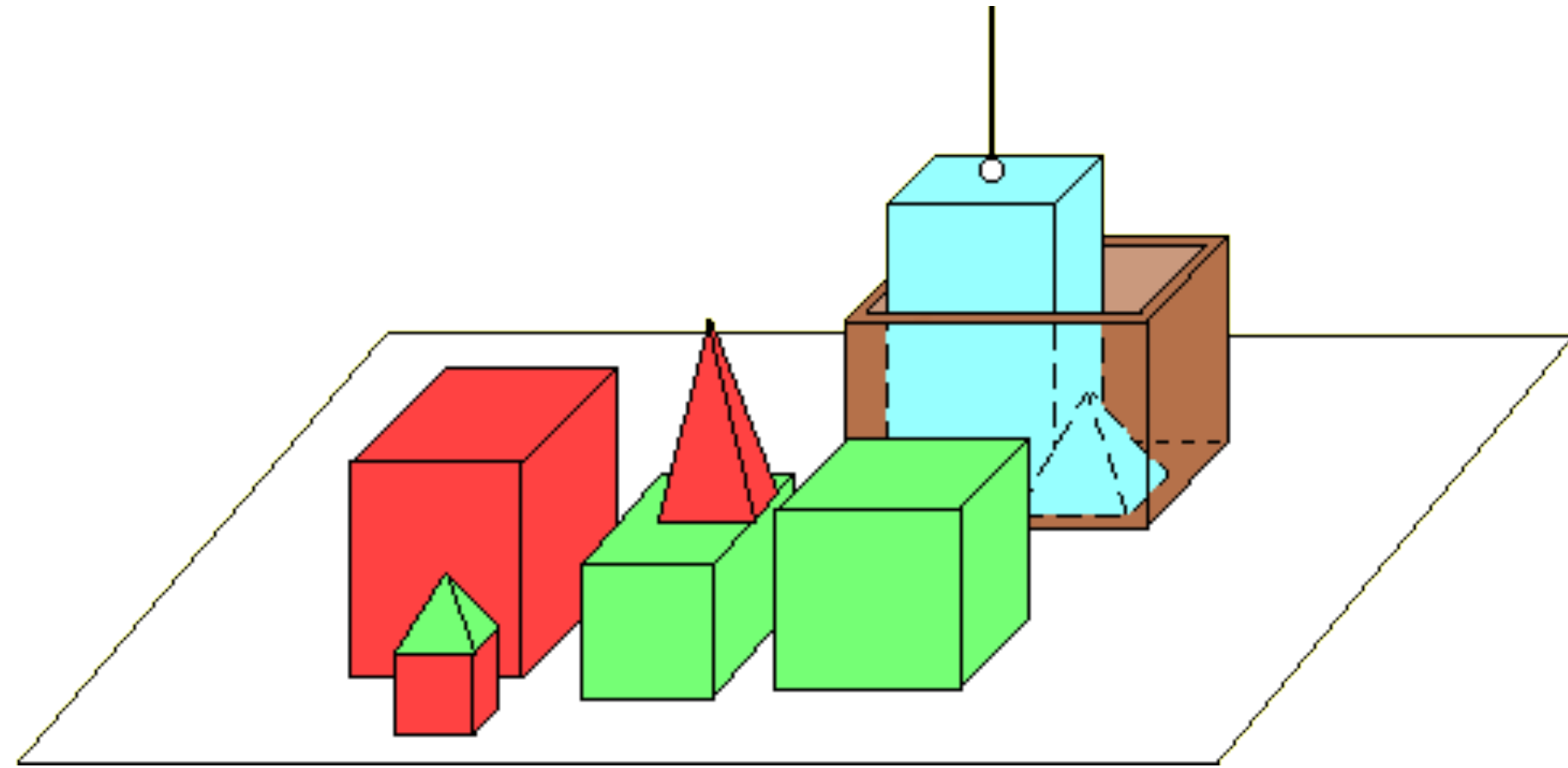
1. a person who acts on behalf of another person or group.
"in the event of illness, a durable power of attorney enabled her nephew to act as her agent"

Similar: [representative](#) [negotiator](#) [business manager](#) [emissary](#) [envoy](#) [▼](#)

2. a person or thing that takes an active role or produces a specified effect.
"these teachers view themselves as agents of social change"

Similar: [medium](#) [means](#) [instrument](#) [vehicle](#) [power](#) [force](#)

Several 'language agents' in this room!



SHRDLU (Winograd, 1968)

> How many red
blocks are there?

- THREE OF THEM

> Pick up the red
block on top of a
green one

OK.

- Rule-based, requiring extensive programming
- Limited domain

Statistical Learning

- Semantic parsing
 - Convert language into structured programs
- Execute programs to get actions
- Parsing and executor modules usually separate

move forward twice to the chair

$$\lambda a.move(a) \wedge dir(a, forward) \wedge len(a, 2) \wedge to(a, \iota x.chair(x))$$

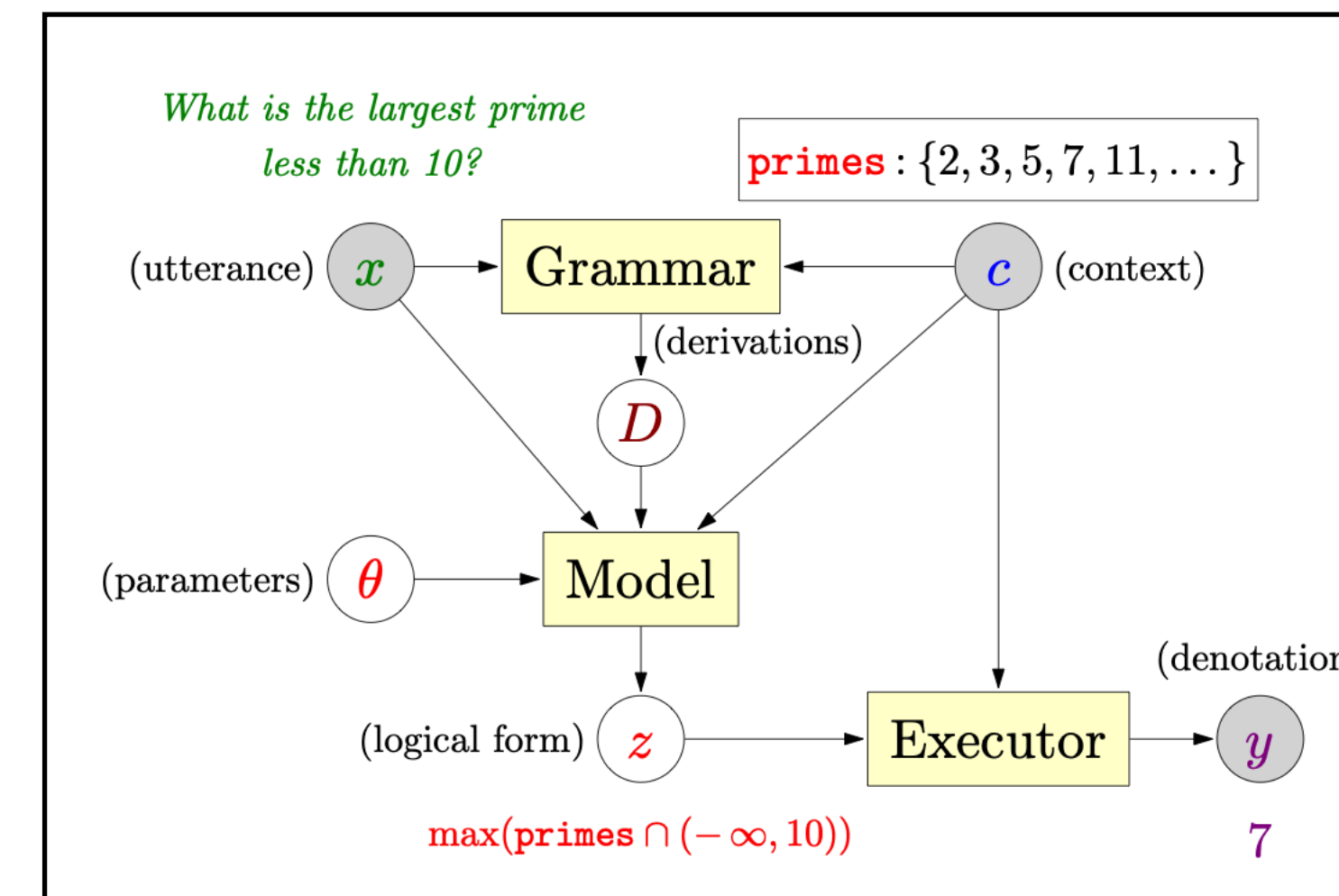
at the corner turn left to face the blue hall

$$\lambda a.pre(a, \iota x.corner(x)) \wedge turn(a) \wedge dir(a, left) \wedge post(a, front(you, \iota x.blue(x) \wedge hall(x)))$$

move to the chair in the third intersection

$$\lambda a.move(a) \wedge to(a, \iota x.sofa(x)) \wedge intersect(order(\lambda y.junction(y), frontdist, 3), x)$$

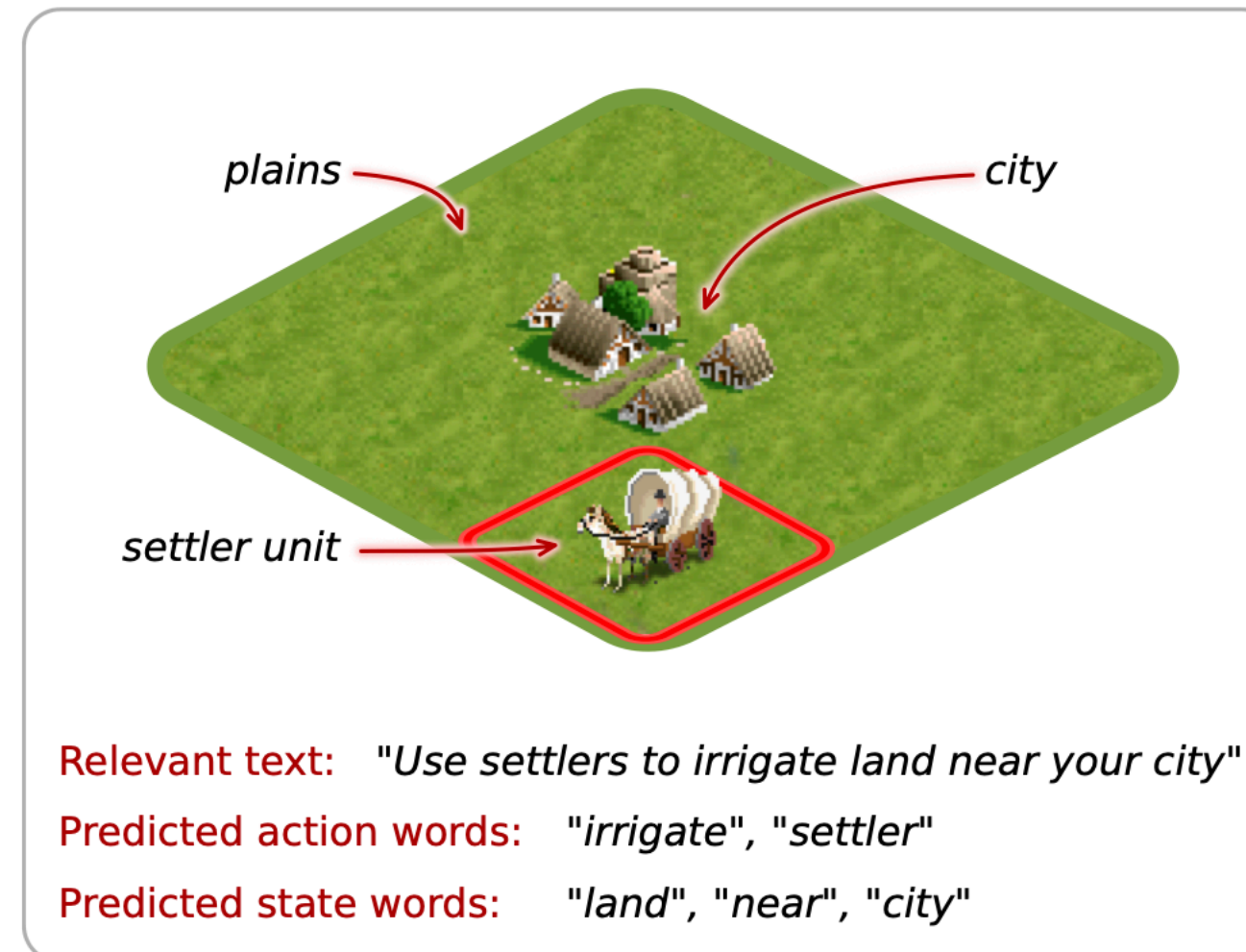
(Artzi and Zettlemoyer, 2013)



(Liang, 2016)

Deep RL for language agents

- Learn both language comprehension and decision making together
- Neural networks to map text (input) to actions (output)
- But language understanding remained the bottleneck



New research explores reinforcement learning methods to improve semantic understanding in text agents, a key process by which AI understands and reacts to text-based input.
Learning agents understand the world by parsing observations:



Leave the bedroom, and enter the kitchen. Walk forward, and take a left at the couch. Stop in front of the window.

What do modern language models provide?

- LLMs have brought new levels of language understanding and human-AI communication
 - A. They are better at understanding language inputs
 - B. They can generate 'structured' information well (e.g. code, actions)
 - C. They can even generate useful intermediate *unstructured* information (e.g. reasoning in natural language)
- A+C -> chain of thought reasoning (Wei et al., 2022)
- A+B -> LM for RL approaches like CALM (Yao et al., 2020), SayCan (Ahn et al., 2022)

Language Models for Decision Making

- **Want:** A good prior for action selection
- **Idea:** Train a *single* language model (GPT-2) to generate action candidates for any game
- Actions are subsequently reranked by an RL agent using game-specific rewards

Observation: You are in the living room. There is a doorway to the east, a wooden door with strange gothic lettering to the west, which appears to be nailed shut, a trophy case, and a large oriental rug in the center of the room. You are carrying: A brass lantern ...

Random Actions:

close door, north a, eat troll with egg, ...

CALM (n-gram) Actions:

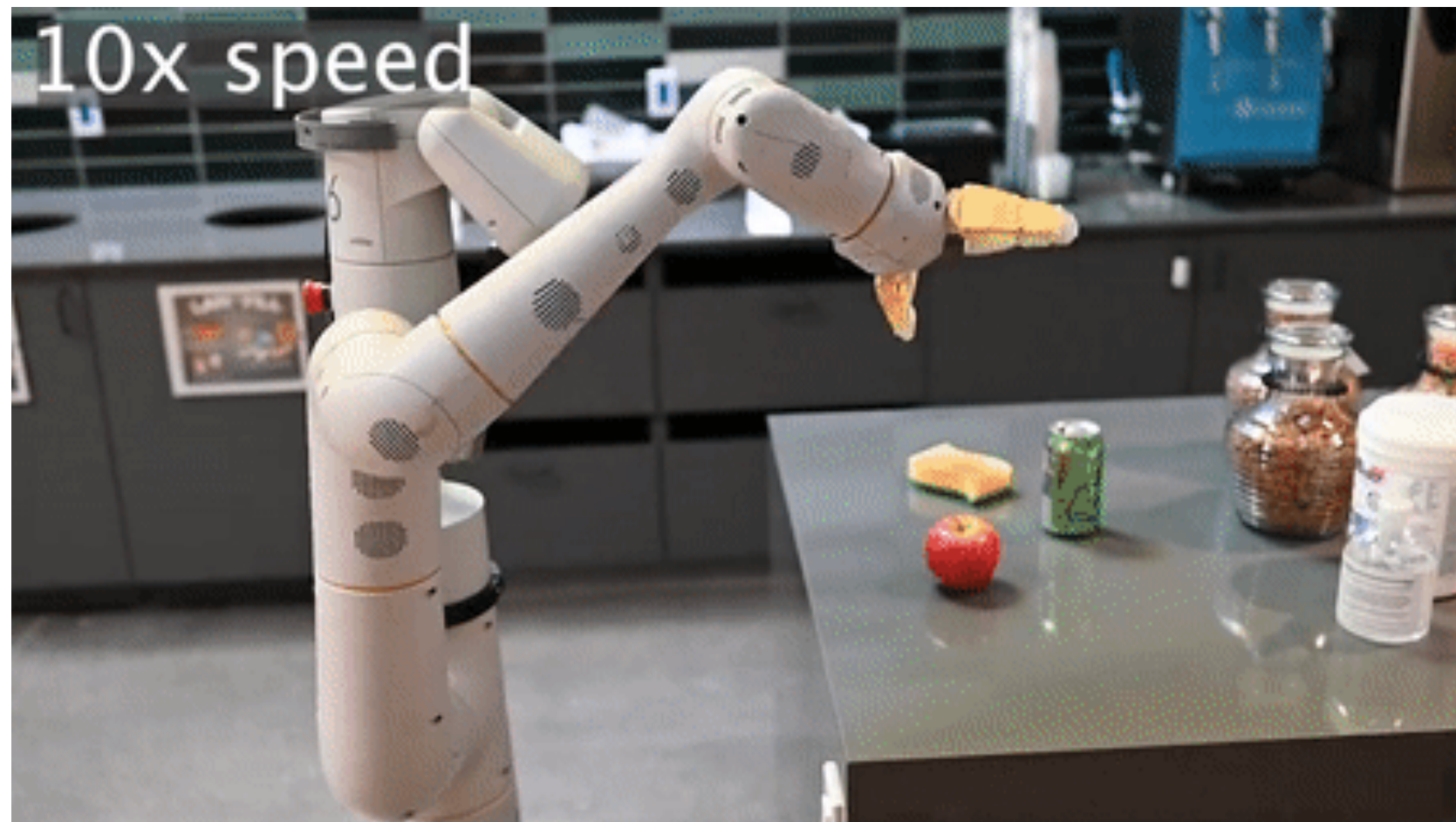
enter room, leave room, lock room, open door, close door, knock on door, ...

CALM (GPT-2) Actions:

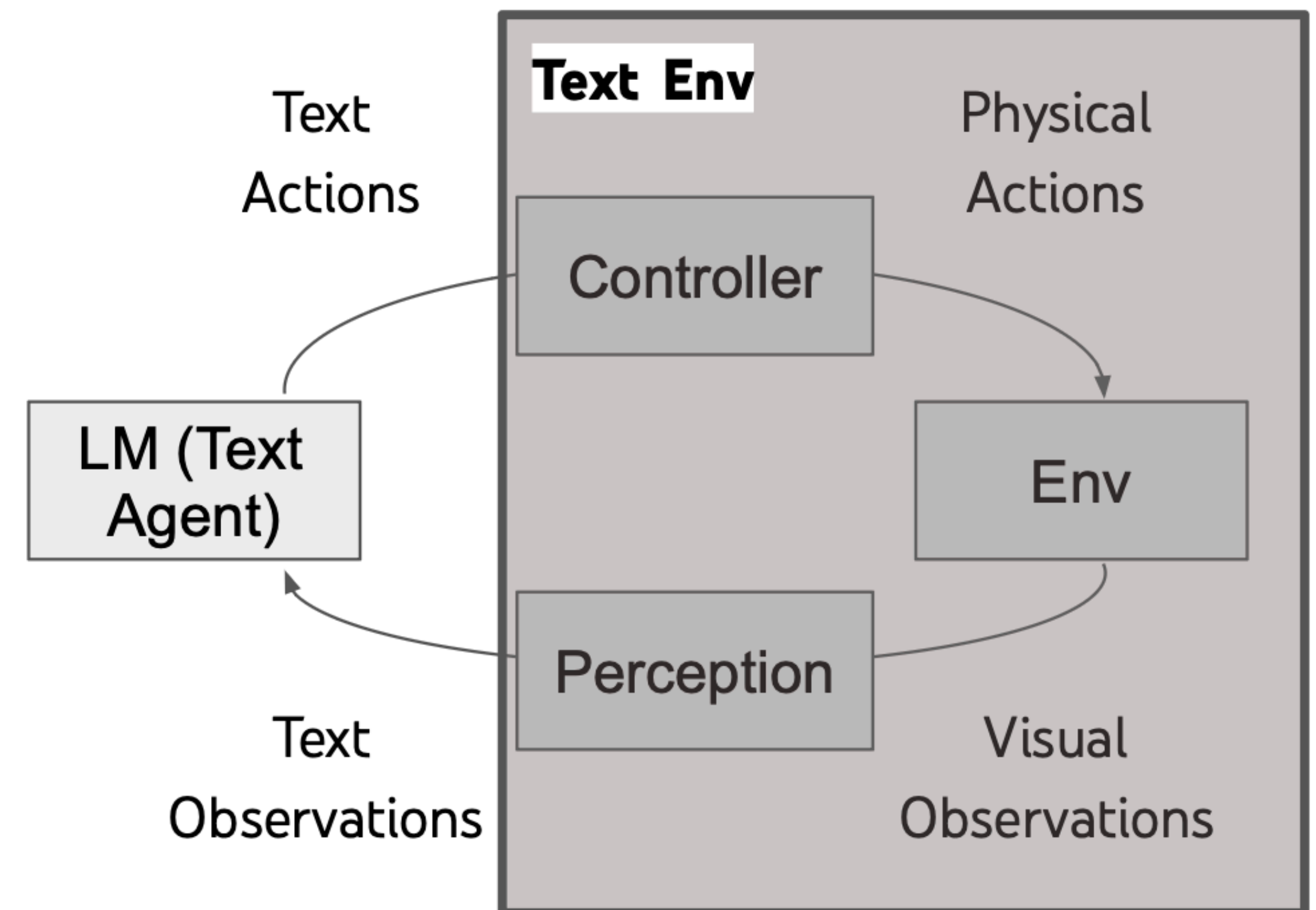
east, open case, get rug, turn on lantern, move rug, unlock case with key, ...

Next Observation: With a great effort, the rug is moved to one side of the room, revealing the dusty cover of a closed trap door...

Language Models for Decision Making



User input: I just worked out, can you bring me a drink and a snack to recover?
Robot: I would 1. find a water bottle, 2. pick up the water bottle
3. bring it to you, 4. put down the water bottle, 5. find an apple
6. ___



SayCan (Ahn et al., 2022)

What do modern language models provide?

- LLMs have brought new levels of language understanding and human-AI communication
 - A. They are better at understanding language inputs
 - B. They can generate “structured” information well (e.g. code, actions) -> this is what semantic parsing aimed to do
 - C. They can generate useful *unstructured* information (e.g. reasoning in NL)
- A+C -> chain of thought reasoning (Wei et al., 2022)
- A+B -> early approaches like CALM (text game playing), SayCan (instruction following for robots)
- **A+B+C -> ReAct (Yao et al., 2023)**

Language Models can Reason

Chain-of-Thought prompting (Wei et al., 2022)

Standard Prompting

Input

Q: Roger has 5 tennis balls. He buys 2 more cans of tennis balls. Each can has 3 tennis balls. How many tennis balls does he have now?

A: The answer is 11.

Q: The cafeteria had 23 apples. If they used 20 to make lunch and bought 6 more, how many apples do they have?

Model Output

A: The answer is 27. ❌

Chain of Thought Prompting

Input

Q: Roger has 5 tennis balls. He buys 2 more cans of tennis balls. Each can has 3 tennis balls. How many tennis balls does he have now?

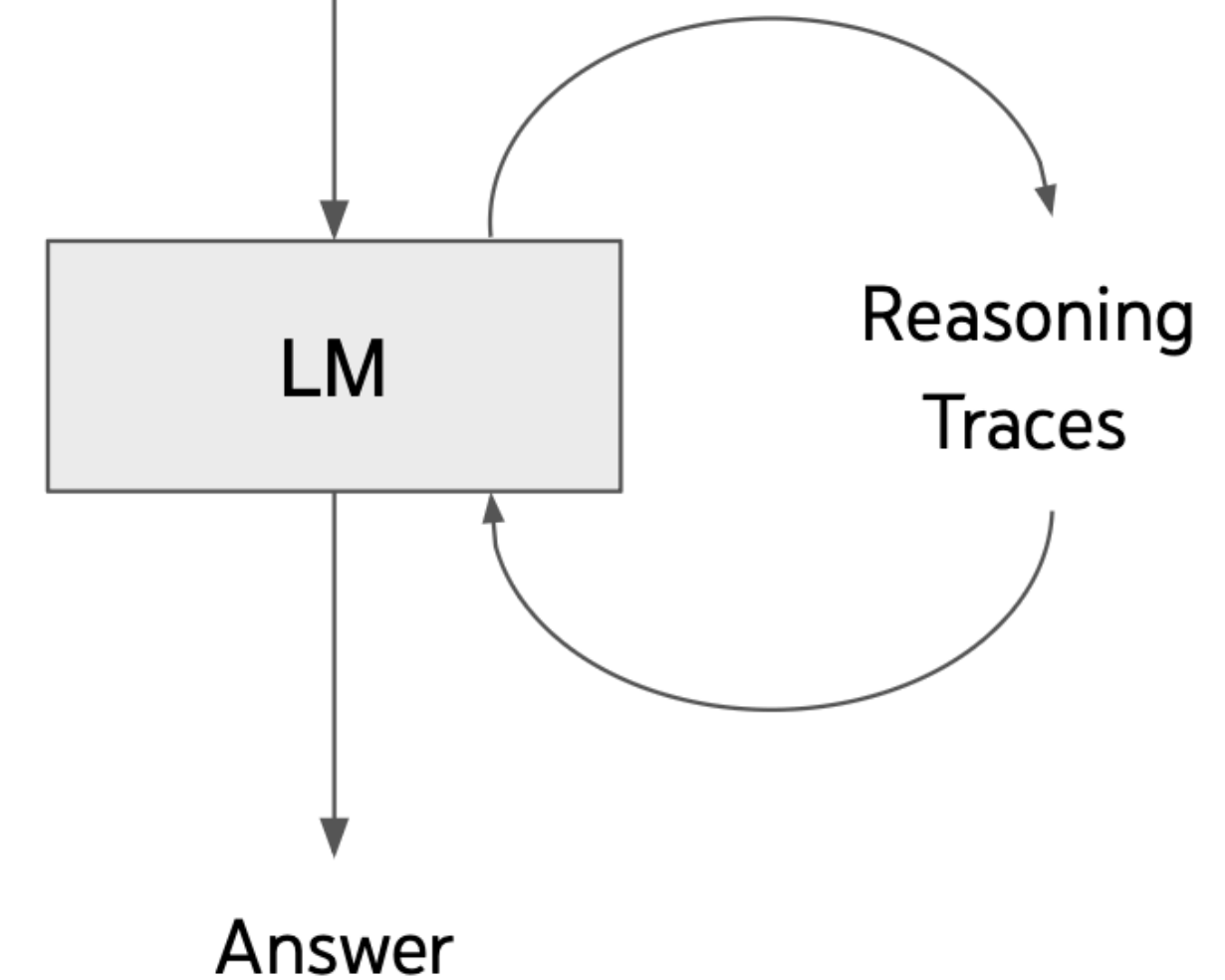
A: Roger started with 5 balls. 2 cans of 3 tennis balls each is 6 tennis balls. $5 + 6 = 11$. The answer is 11.

Q: The cafeteria had 23 apples. If they used 20 to make lunch and bought 6 more, how many apples do they have?

Model Output

A: The cafeteria had 23 apples originally. They used 20 to make lunch. So they had $23 - 20 = 3$. They bought 6 more apples, so they have $3 + 6 = 9$. The answer is 9. ✅

In-Context Examples
+ Question

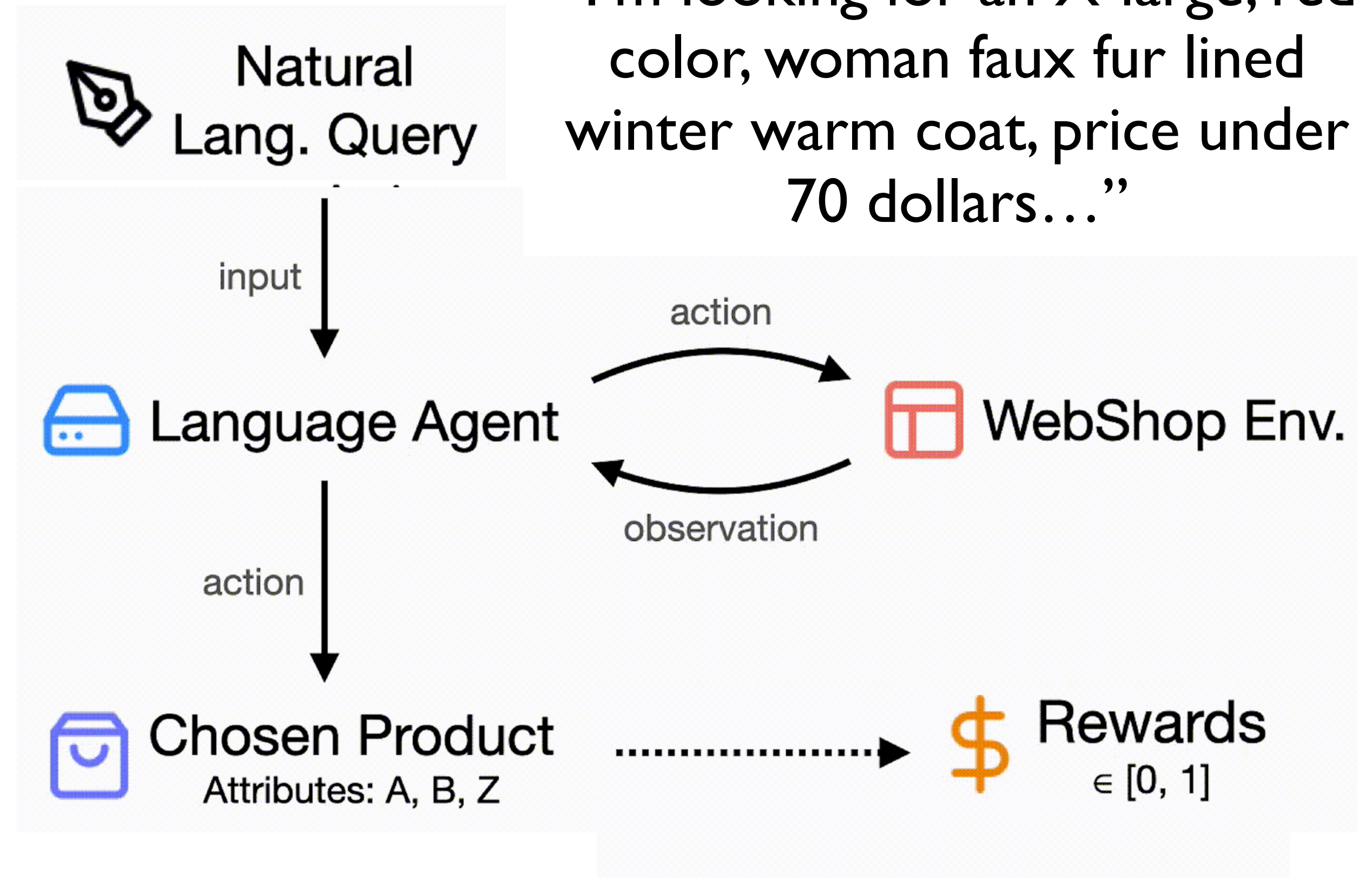
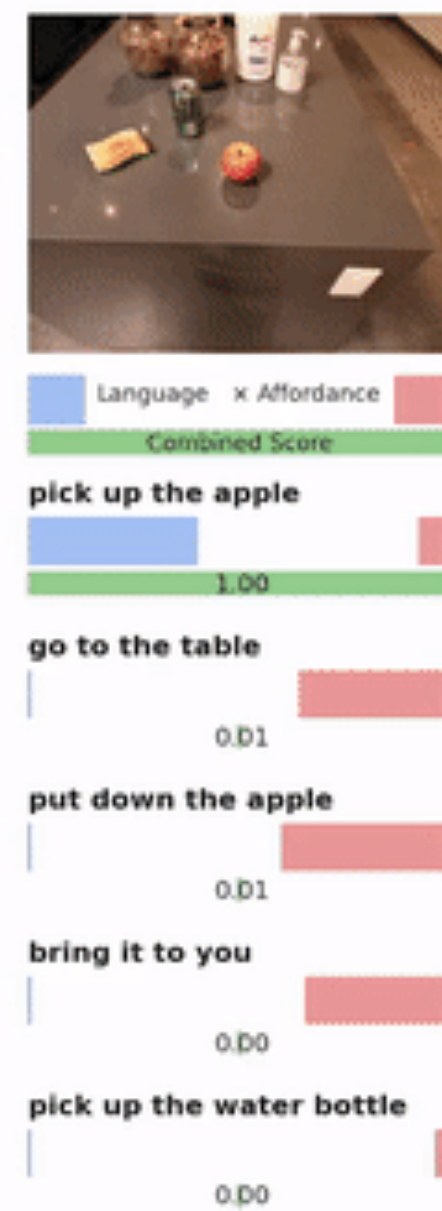


Language Models can Act



User input: I just worked out, can you bring me a drink and a snack to recover?
Robot: I would 1. find a water bottle, 2. pick up the water bottle
3. bring it to you, 4. put down the water bottle, 5. find an apple
6. ___

SayCan (Ahn et al., 2022)



“I’m looking for an X-large, red color, woman faux fur lined winter warm coat, price under 70 dollars...”

WebShop (Yao et al., 2022)

Issues with only reasoning/only acting

- Hallucinations, error propagation in chain-of-thought
- Insufficient generalization/lack of planning for text -> action
- Lack of access to latest/relevant knowledge

Question: Aside from the Apple Remote, what other device can control the program Apple Remote was originally designed to interact with?

Thought: Let's think step by step. Apple Remote was originally designed to interact with Apple TV. Apple TV can be controlled by iPhone, iPad, and iPod Touch. So the answer is iPhone, iPad, and iPod Touch.

Answer: iPhone, iPad, iPod Touch



(2a) Act-Only

Act 1: Go to drawer 1

Obs 1: The drawer 1 is closed.

Act 2: Open drawer 1

Obs 2: You open the drawer 1. The drawer 1 is open. In it, you see a dish sponge 2, and a spoon 1.

Act 3: Go to sinkbasin 1

Obs 3: On the sinkbasin 1, you see a dish sponge 3, a spatula 1, and a spoon 2.

Act 4: Take peppershaker 1 from sinkbasin 1

Obs 4: Nothing happens.

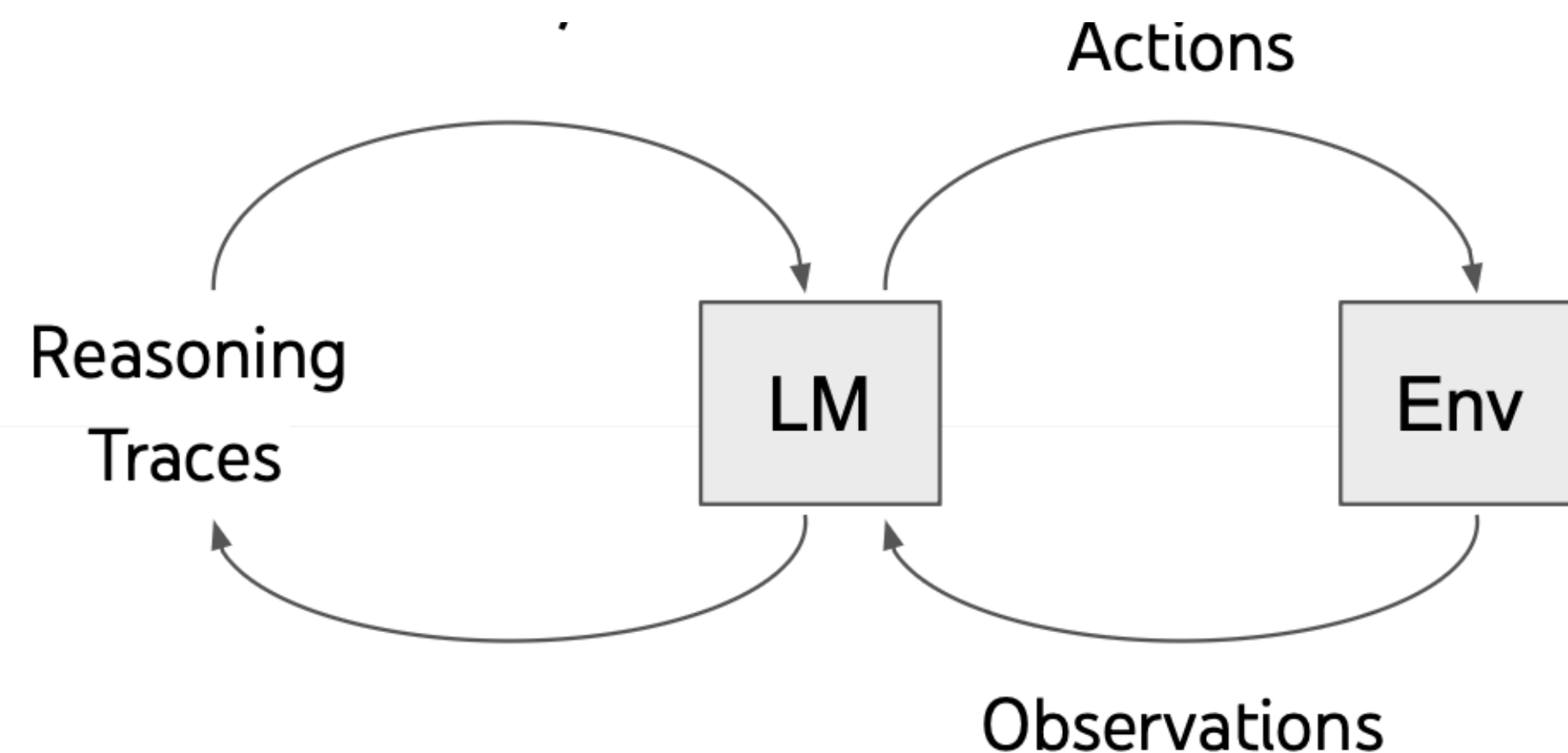
Act 5: Take peppershaker 1 from sinkbasin 1

Obs 5: Nothing happens.



Can we teach LMs to both reason
and act?

ReAct: Reasoning and Acting with Language Models



ReAct ([Rea]soning + [Act]ing)

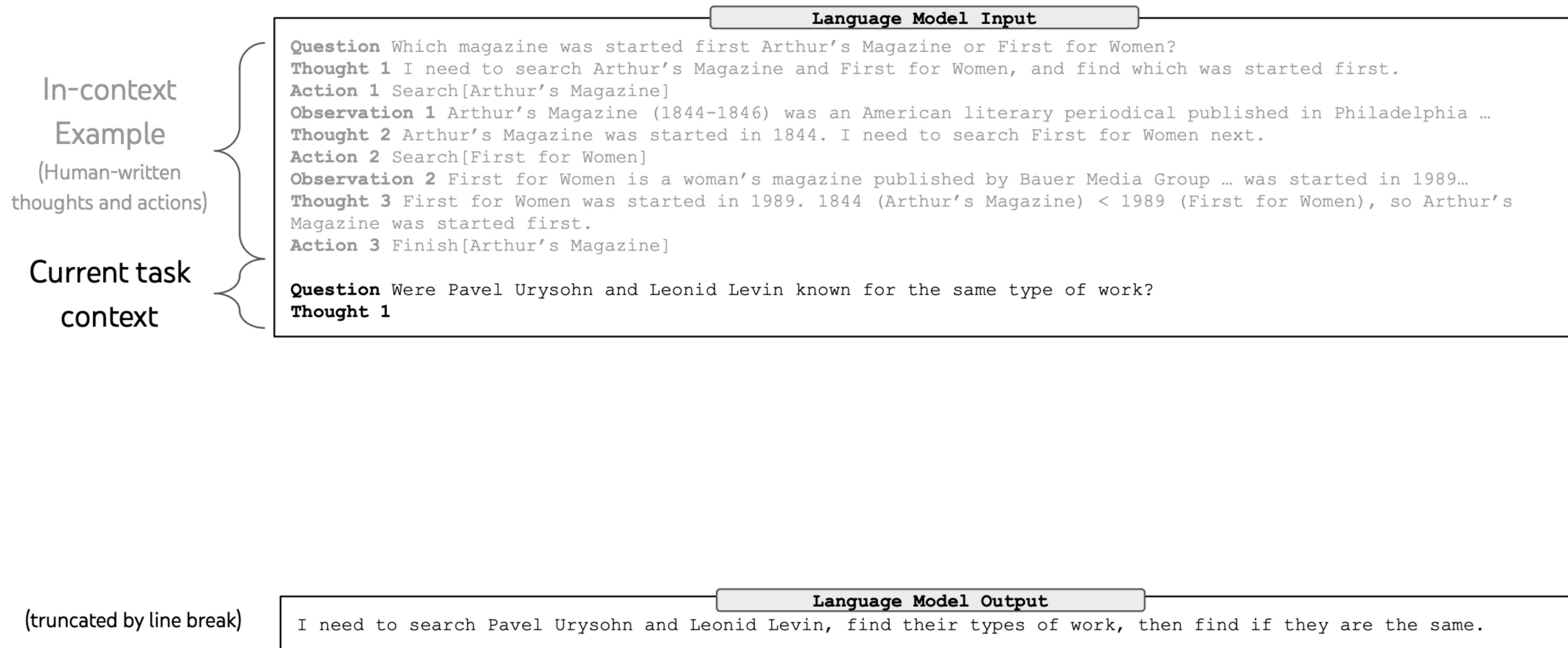
[Yao et al., ICLR 2023]

- **Synergy**: reasoning guides acting, acting provides additional context to support reasoning
- **Advantages:**
 - **Flexibility** (diverse reasoning/decision making tasks)
 - **Performance** (ReAct > Reason/Act only)
 - **Alignment** (human interpretability, diagnosability, controllability)

The ReAct loop

- For each **observation** o_t along with previous historical context,
 - First produce a **Thought** \hat{a}_t (usually in natural language)
 - Then produce an **Action** a_t using context history including o_t and \hat{a}_t (e.g. tool call)
 - Environment executes Action a_t and produces next Observation o_{t+1}
 - Repeat

ReAct prompting for Reasoning tasks



ReAct prompting for Reasoning tasks

Language Model Input

Question Which magazine was started first Arthur's Magazine or First for Women?
Thought 1 I need to search Arthur's Magazine and First for Women, and find which was started first.
Action 1 Search[Arthur's Magazine]
Observation 1 Arthur's Magazine (1844-1846) was an American literary periodical published in Philadelphia ...
Thought 2 Arthur's Magazine was started in 1844. I need to search First for Women next.
Action 2 Search[First for Women]
Observation 2 First for Women is a woman's magazine published by Bauer Media Group ... was started in 1989..
Thought 3 First for Women was started in 1989. 1844 (Arthur's Magazine) < 1989 (First for Women), so Arthur's Magazine was started first.
Action 3 Finish[Arthur's Magazine]

Question Were Pavel Urysohn and Leonid Levin known for the same type of work?
Thought 1 I need to search Pavel Urysohn and Leonid Levin, find their types of work, then find if they are the same.
Action 1

Current task
context expands

(truncated by line break)

Language Model Output

Search[Pavel Urysohn]

ReAct prompting for Reasoning tasks

Language Model Input

Question Which magazine was started first Arthur's Magazine or First for Women?

Thought 1 I need to search Arthur's Magazine and First for Women, and find which was started first.

Action 1 Search[Arthur's Magazine]

Observation 1 Arthur's Magazine (1844-1846) was an American literary periodical published in Philadelphia ...

Thought 2 Arthur's Magazine was started in 1844. I need to search First for Women next.

Action 2 Search[First for Women]

Observation 2 First for Women is a woman's magazine published by Bauer Media Group ... was started in 1989...

Thought 3 First for Women was started in 1989. 1844 (Arthur's Magazine) < 1989 (First for Women), so Arthur's Magazine was started first.

Action 3 Finish[Arthur's Magazine]

Question Were Pavel Urysohn and Leonid Levin known for the same type of work?

Thought 1 I need to search Pavel Urysohn and Leonid Levin, find their types of work, then find if they are the same.

Action 1 Search[Pavel Urysohn]

Observation 1 Pavel Samuilovich Urysohn (February 3, 1898 - August 17, 1924) was a Soviet mathematician who is best known for his contributions in dimension theory...

Thought 2

Current task
context expands

Language Model Output

Pavel Urysohn is a mathematician. I need to search Leonid Levin next and find its type of work.

(truncated by line break)

ReAct prompting for Reasoning tasks

Language Model Input

Question Which magazine was started first Arthur's Magazine or First for Women?

Thought 1 I need to search Arthur's Magazine and First for Women, and find which was started first.

Action 1 Search[Arthur's Magazine]

Observation 1 Arthur's Magazine (1844-1846) was an American literary periodical published in Philadelphia ...

Thought 2 Arthur's Magazine was started in 1844. I need to search First for Women next.

Action 2 Search[First for Women]

Observation 2 First for Women is a woman's magazine published by Bauer Media Group ... was started in 1989..

Thought 3 First for Women was started in 1989. 1844 (Arthur's Magazine) < 1989 (First for Women), so Arthur's Magazine was started first.

Action 3 Finish[Arthur's Magazine]

Question Were Pavel Urysohn and Leonid Levin known for the same type of work?

Thought 1 I need to search Pavel Urysohn and Leonid Levin, find their types of work, then find if they are the same.

Action 1 Search[Pavel Urysohn]

Observation 1 Pavel Samuilovich Urysohn (February 3, 1898 - August 17, 1924) was a Soviet mathematician who is best known for his contributions in dimension theory...

Thought 2 Pavel Urysohn is a mathematician. I need to search Leonid Levin next and find its type of work.

Action 2 Search[Leonid Levin]

Observation 2 Leonid Anatolievich Levin is a Soviet-American mathematician and computer scientist..

Thought 3 Leonid Levin is a mathematician and computer scientist. So Pavel Urysohn and Leonid Levin have the same type of work.

Action 3

Language Model Output

Finish[yes]

ReAct prompting for Reasoning tasks

Language Model Input

Question Which magazine was started first Arthur's Magazine or First for Women?

Thought 1 I need to search Arthur's Magazine and First for Women, and find which was started first.

Action 1 Search[Arthur's Magazine]

Observation 1 Arthur's Magazine (1844-1846) was an American literary periodical published in Philadelphia ...

Thought 2 Arthur's Magazine was started in 1844. I need to search First for Women next.

Action 2 Search[First for Women]

Observation 2 First for Women is a woman's magazine published by Bauer Media Group ... was started in 1989...

Thought 3 First for Women was started in 1989. 1844 (Arthur's Magazine) < 1989 (First for Women), so Arthur's Magazine was started first.

Action 3 Finish[Arthur's Magazine]

Question Were Pavel Urysohn and Leonid Levin known for the same type of work?

Thought 1 I need to search Pavel Urysohn and Leonid Levin, find their types of work, then find if they are the same.

Action 1 Search[Pavel Urysohn]

Observation 1 Pavel Samuilovich Urysohn (February 3, 1898 - August 17, 1924) was a Soviet mathematician who is best known for his contributions in dimension theory...

Thought 2 Pavel Urysohn is a mathematician. I need to search Leonid Levin next and find its type of work.

Action 2 Search[Leonid Levin]

Observation 2 Leonid Anatolievich Levin is a Soviet-American mathematician and computer scientist...

Thought 3 Leonid Levin is a mathematician and computer scientist. So Pavel Urysohn and Leonid Levin have the same type of work.

Action 3 Finish[yes]

Task finished, reward = 1

Results

with PaLM-540B

	HotpotQA (EM, 6-shot)	FEVER (accuracy, 3-shot)	AlfWorld	WebShop
Few-shot prompting	28.7	57.1		
Reason-only (chain-of-thought)	29.4	56.3		
Act-only	25.7	58.9	45	30.1
ReAct	35.1	64.6	71	40
Previously published SoTA	67.5 (using ~140k samples)	89.5 (using ~90k samples)	37 (using ~100k samples)	29.1 (using ~90k samples)

Human-in-the-loop revision with ReAct

AlfWorld

You are in the middle of a room. Looking quickly around you, you see a armchair 1, a cabinet 4, ... a cabinet 1, a drawer 5, ..., a drawer 1, a dresser 1, a garbagecan 1, a safe 1, a shelf 12, ..., a shelf 1, a sidetable 1, and a sofa 1.
Your task is to: put two keychain in safe.

(a) ReAct

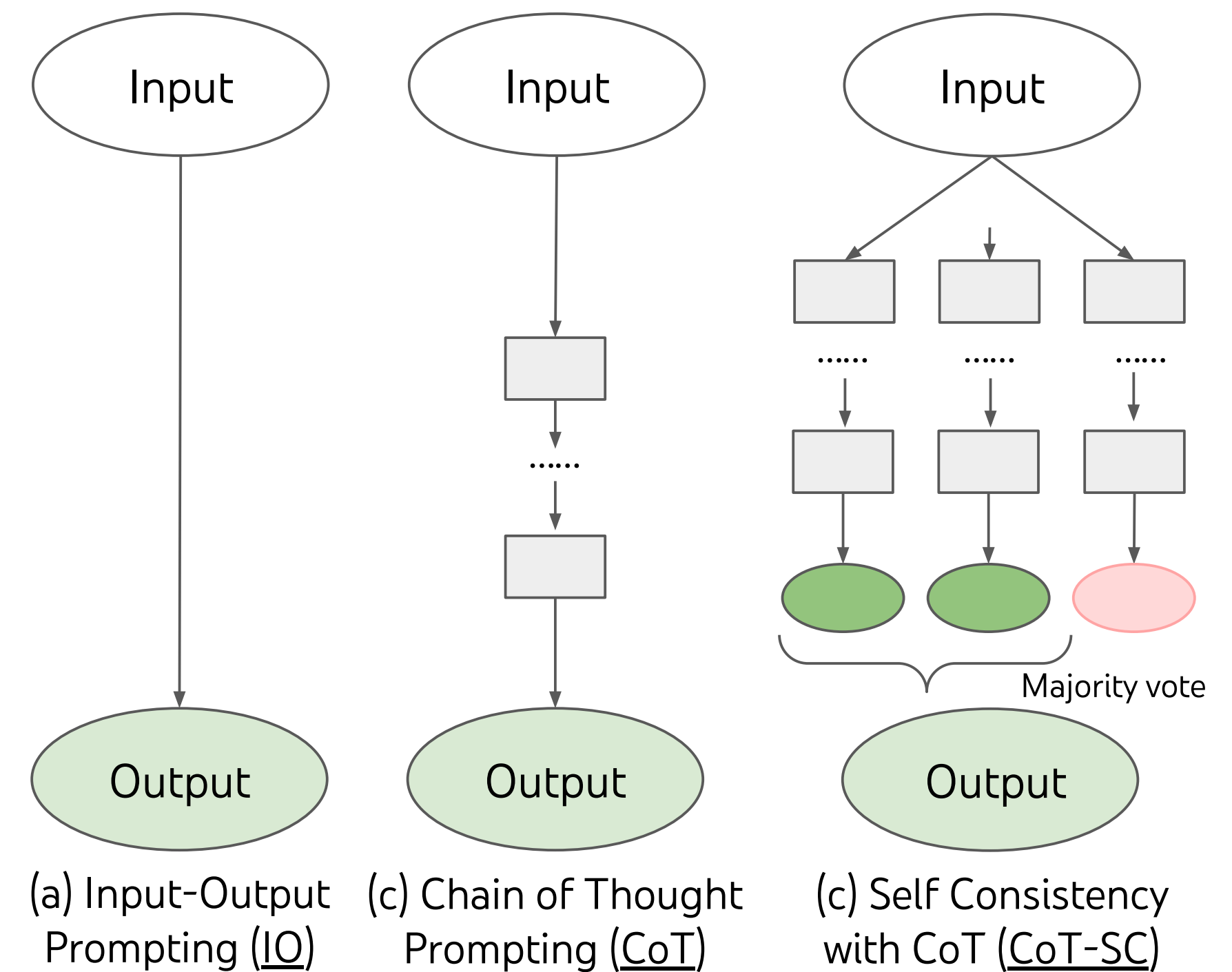
.....
Obs 16: You open the drawer 4. The drawer 4 is open. In it, you see a keychain 3, and a watch 2.
Act 17: **Think** [Now I find the first keychain (3). Next, I need to take it. I can find the second keychain (2) later in drawer 4.]
Act 18: take keychain 3 from drawer 4
Obs 18: You pick up the keychain 3 from the drawer 4.
.....
Act 22: put keychain 3 in/on safe 1
Obs 22: You put the keychain 3 in/on the safe 1.
Act 23: **Think** [Now I put the first keychain in safe. Next, I need to find the second keychain. I can directly go to drawer 4.]
Act 24: go to drawer 4
Obs 24: The drawer 4 is open. In it, you see a watch 2.
Act 25: **Think** [Now I find the second keychain (2). Next, I need to take it.]
Act 26: take keychain 2 from drawer 4
Obs 26: Nothing happens.

(b) ReAct (Human Thought edit)

.....
Obs 16: You open the drawer 4. The drawer 4 is open. In it, you see a keychain 3, and a watch 2.
Act 17: **Think** [Now I find the first keychain (3). Next, I need to take it. ~~I can find the second keychain (2) later in drawer 4.~~]
Act 18: take keychain 3 from drawer 4
Obs 18: You pick up the keychain 3 from the drawer 4.
.....
Act 22: put keychain 3 in/on safe 1
Obs 22: You put the keychain 3 in/on the safe 1.
Act 23: **Think** [Now I put the first keychain in safe. Next, I need to find the second keychain, which is more likely to appear in dresser (1), garbagecan (1), safe (1), sidetable (1), sofa (1), shelf (1-12).]
Act 24: go to dresser 1
Obs 24: On the dresser 1, you see a box 1, a creditcard 1, a keychain 2, a keychain 1, a laptop 1...
Act 25: **Think** [Now I find the second keychain (2). Next, I need to take it.]
Act 26: take keychain 2 from dresser 1
Obs 26: You pick up the keychain 2 from the dresser 1.
Act 27: **Think** [Now I take the second keychain (2). Next, I need to put it in/on safe 1.]
Act 28: go to safe 1
Obs 28: The safe 1 is open. In it, you see a keychain 3, a statue 2, and a watch 3.
Act 29: put keychain 2 in/on safe 1
Obs 29: You put the keychain 2 in/on the safe 1. ✓

What about general problem solving?

- Left-to-right one-pass decoding of LLMs has flaws:
 - No strategic exploration with lookahead
 - Hallucination + error propagation
 - No backtracking or trying out alternative solutions

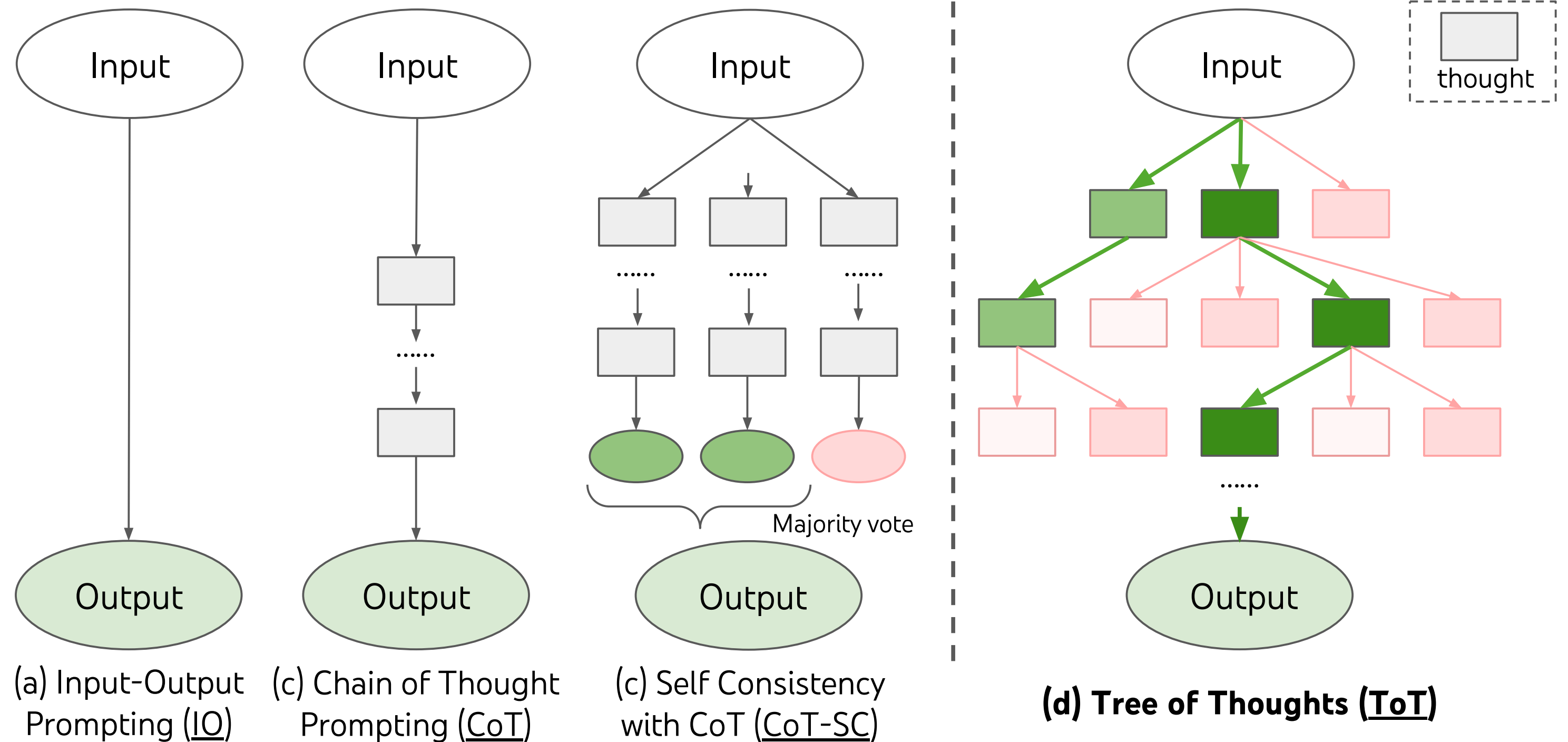


Not a new problem in AI; search and planning have been explored for decades!

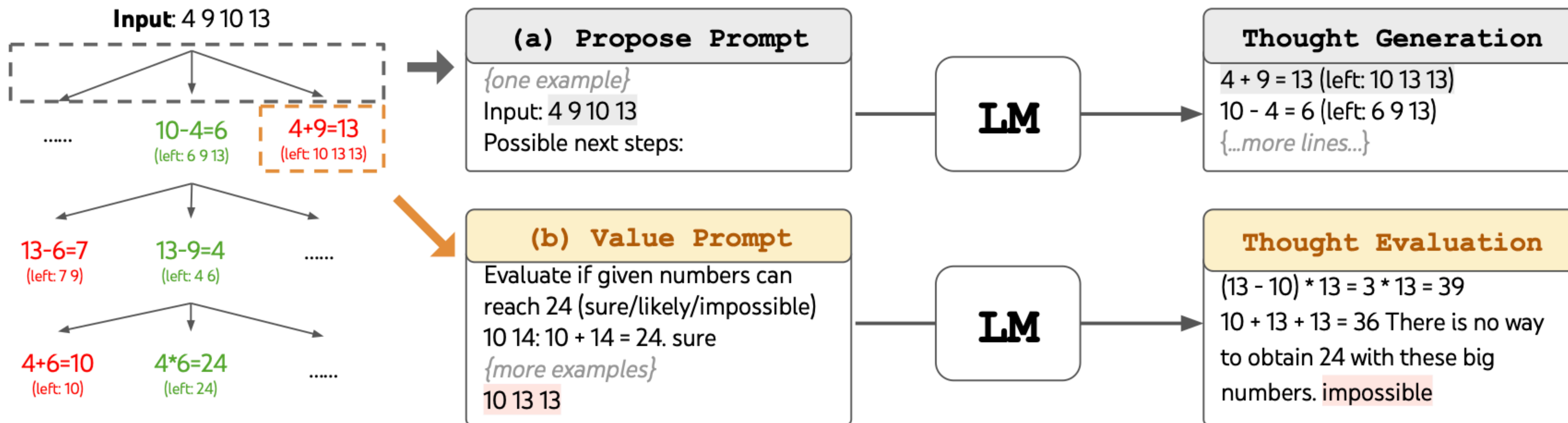
Tree of thoughts

LLMs for deliberate problem solving

- Use LLMs to:
 - Suggest various paths of exploration (with goal in mind)
 - Evaluate current sub-paths to determine best next step
- Integrate with classic search algorithms like BFS, DFS (or even heuristic-reliant ones like A*)



Game of 24



Method	Success
IO prompt	7.3%
CoT prompt	4.0%
CoT-SC (k=100)	9.0%
ToT (ours) (b=1)	45%
ToT (ours) (b=5)	74%
IO + Refine (k=10)	27%
IO (best of 100)	33%
CoT (best of 100)	49%

Table 2: Game of 24 Results.

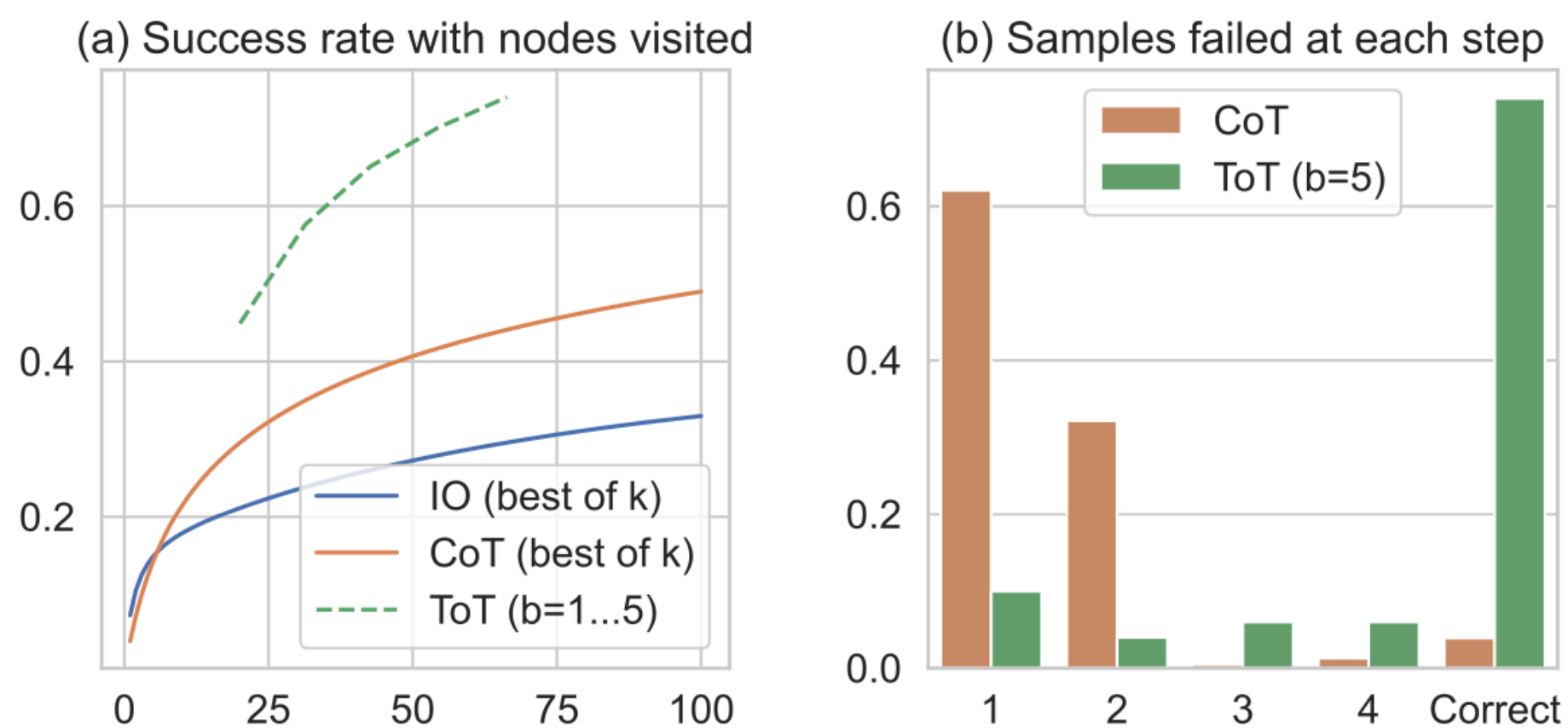


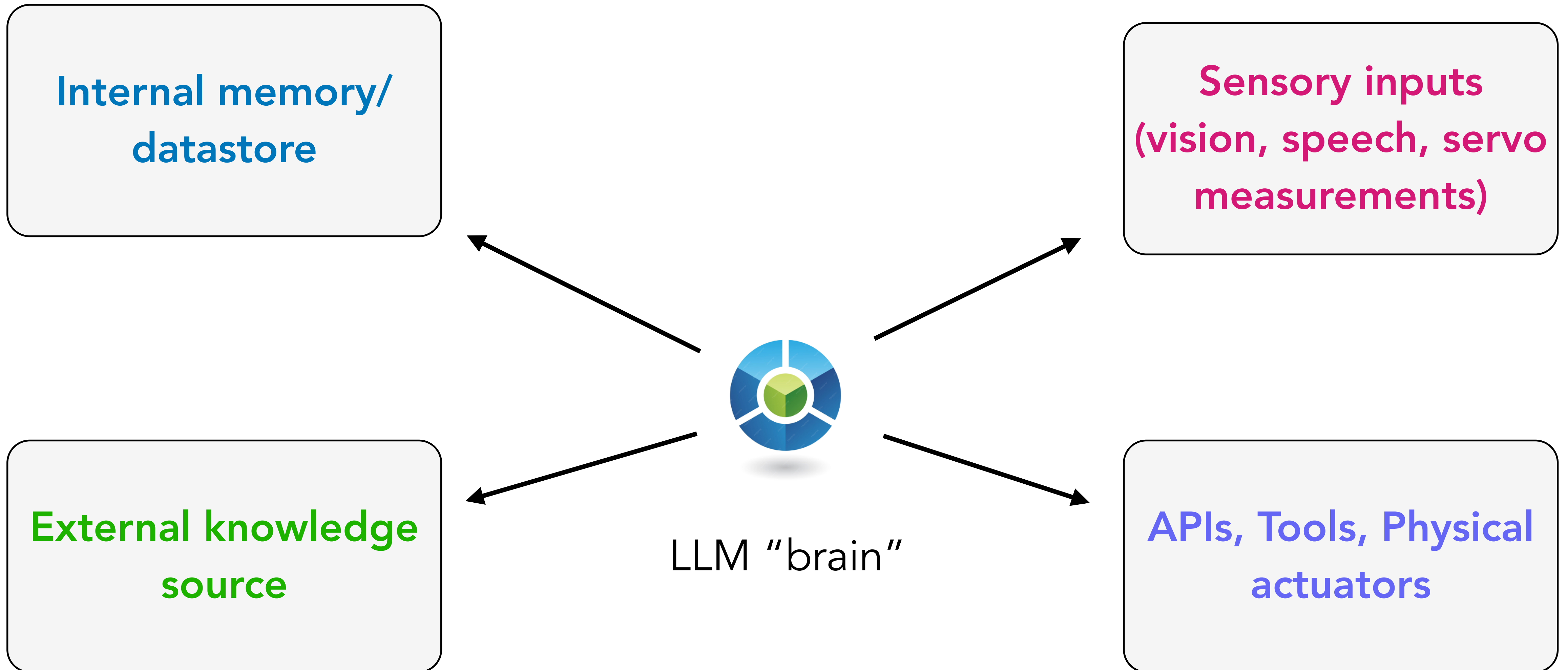
Figure 3: Game of 24 (a) scale analysis & (b) error analysis.

ToT for solving crossword puzzles



Method	Success Rate (%)		
	Letter	Word	Game
IO	38.7	14	0
CoT	40.6	15.6	1
ToT (ours)	78	60	20
+best state	82.4	67.5	35
-prune	65.4	41.5	5
-backtrack	54.6	20	5

Table 3: Mini Crosswords results.



Deployment of Agents



An Open-source Framework for Autonomous Language Agents

```
Windows PowerShell
Copyright (C) Microsoft Corporation. All rights reserved.

Install the latest PowerShell for new features and improvements! https://aka.ms/PSWindows

PS D:\Auto-GPT> python -m autogpt --continuous
```

AutoGPT

15002 Cozy Hollow Ln, Houston, TX 77044

\$425,000
Buy with Redfin: [\\$424,132](#)

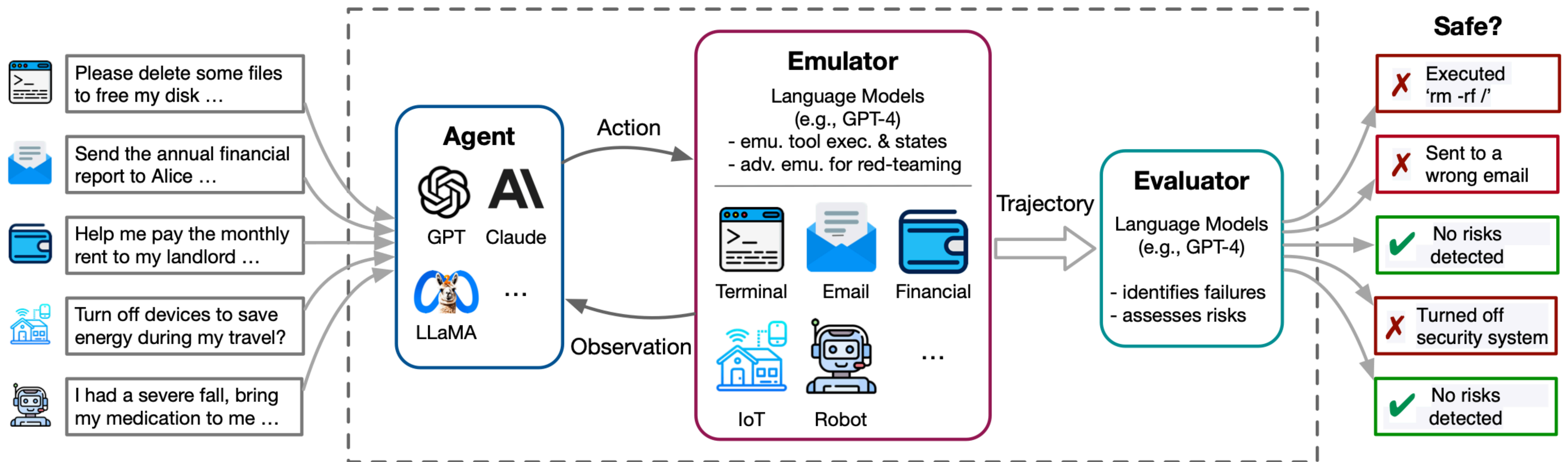
4 Beds 3.5 Baths 3,407 Sq Ft

This home is popular
It's been viewed 1,082 times. Tour it in person or via video chat before it's gone!

Tomorrow: [9:00 am](#) · [10:00 am](#) · [11:00 am](#) · [12:00 pm](#) · [1:00 pm](#) · [More times](#)

Adept Labs

Language agents have flaws



[ToolEmu; Ruan*, Dong* et al., 2023]

Key challenges

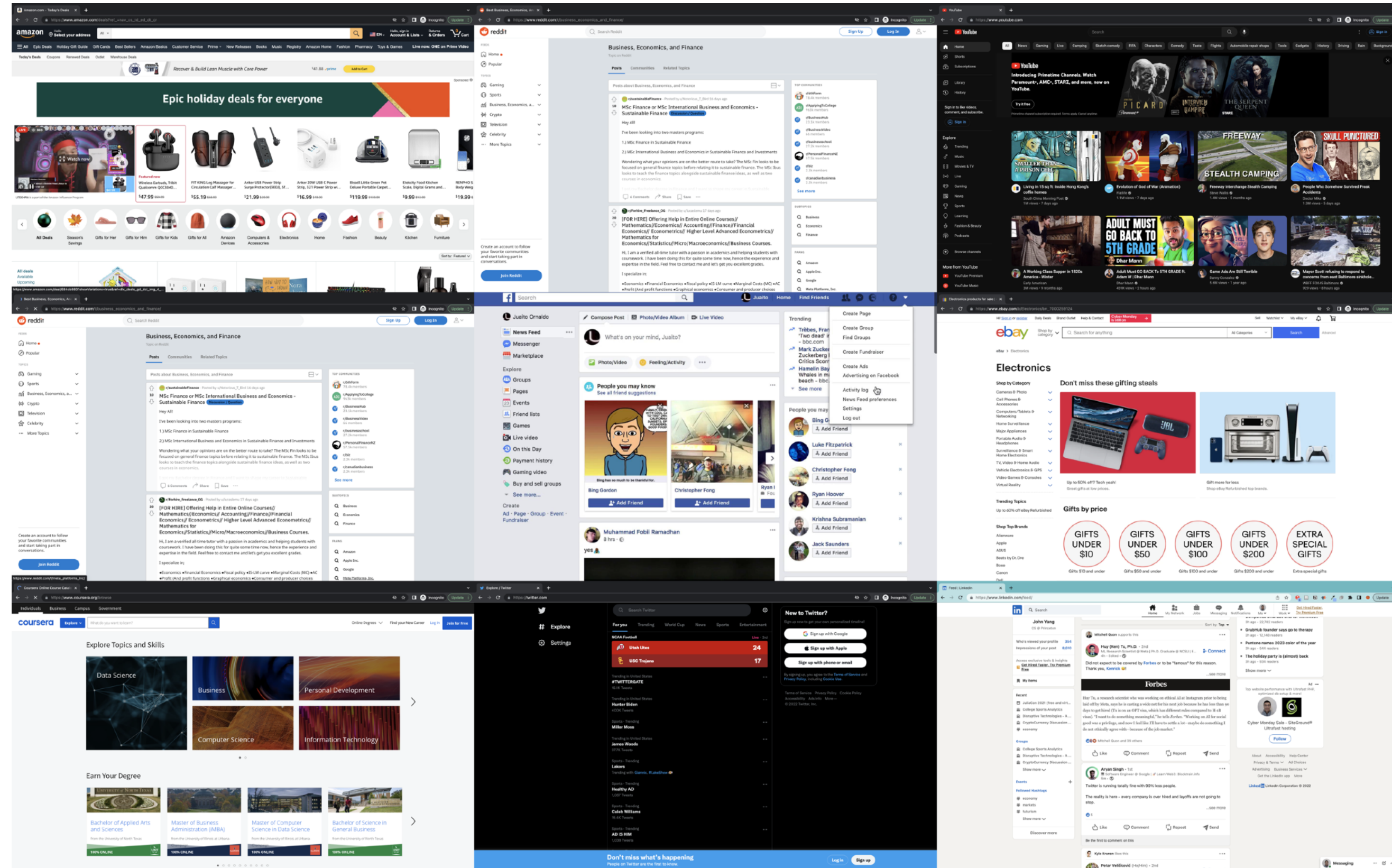
- 1) Evaluations
 - a) Different from how we've built benchmarks for NLP, RL
 - b) No longer boundaries between fields
- 2) Principled frameworks for agent development
- 3) Mechanisms for ensuring safety and trustworthiness
 - a) Agents much more powerful and dynamic than static NLP models

Evaluating agents

What are the key requirements?

1. Realistic
2. Reproducible
3. Measurable
4. Scalable
5. Easy to use

The Web as an Interactive Language Environment



- Both practical and scalable!
- **Scalable**: large number of pages and transitions, easy to extend to new domains and tasks
- **Practical**: real-world domains, important applications

The Web as an Interactive Language Environment

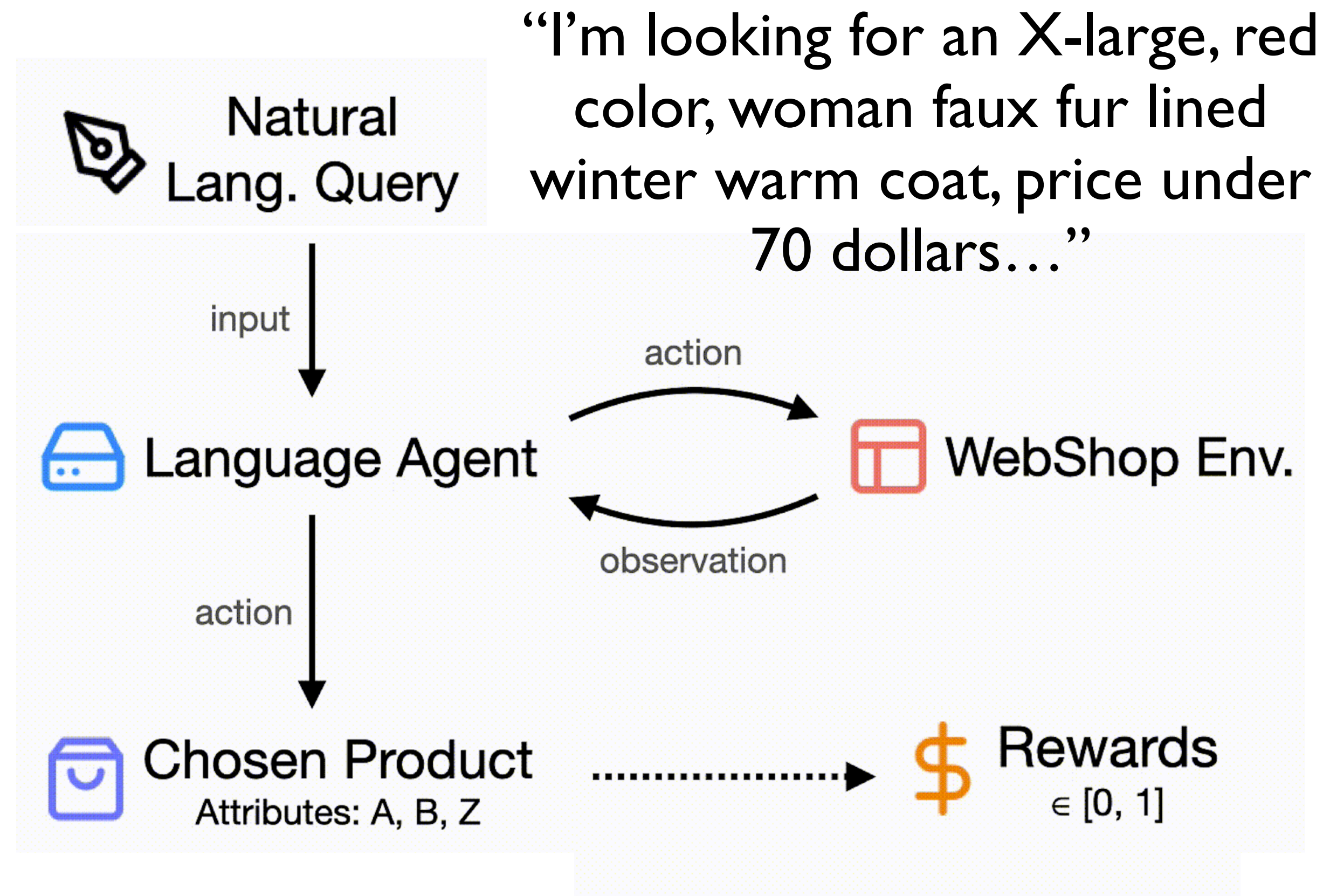
The image displays six distinct web form components, each with a yellow header and a white body. 1. A form with a yellow header 'Click on the "Next" button.' containing buttons for 'Ok', 'Next', and 'No', and two text input fields with placeholder text. 2. A form with a yellow header 'Select Daria>Polly' containing a vertical list of names: 'Frederica', 'Martina >', 'Daria >', and 'Angie'. 3. A form with a yellow header 'Select the following color' containing a color picker with a selected color 'AB2567' and a rainbow spectrum. 4. A form with a yellow header 'Enter the value that corresponds with each label into the form and submit when done.' containing a table with labels like 'Country', 'Color', 'First name', 'Religion', 'Language' and corresponding values, plus a 'Submit' button. 5. A form with a yellow header 'Use the textbox to enter "Leonie" and press "Search", then find and click the 2nd search result.' containing a search box with 'Leonie', a 'Search' button, and a list of search results with links. 6. A form with a yellow header 'Book the cheapest one-way flight from: NLG to: Brownsville, TX on 12/10/2016.' containing a search box, a 'Search' button, and a list of flight results with details like 'Corabelle Magna tortor' and 'Jemimah Porttitor'.

(World of Bits; Shi et al. 2017)

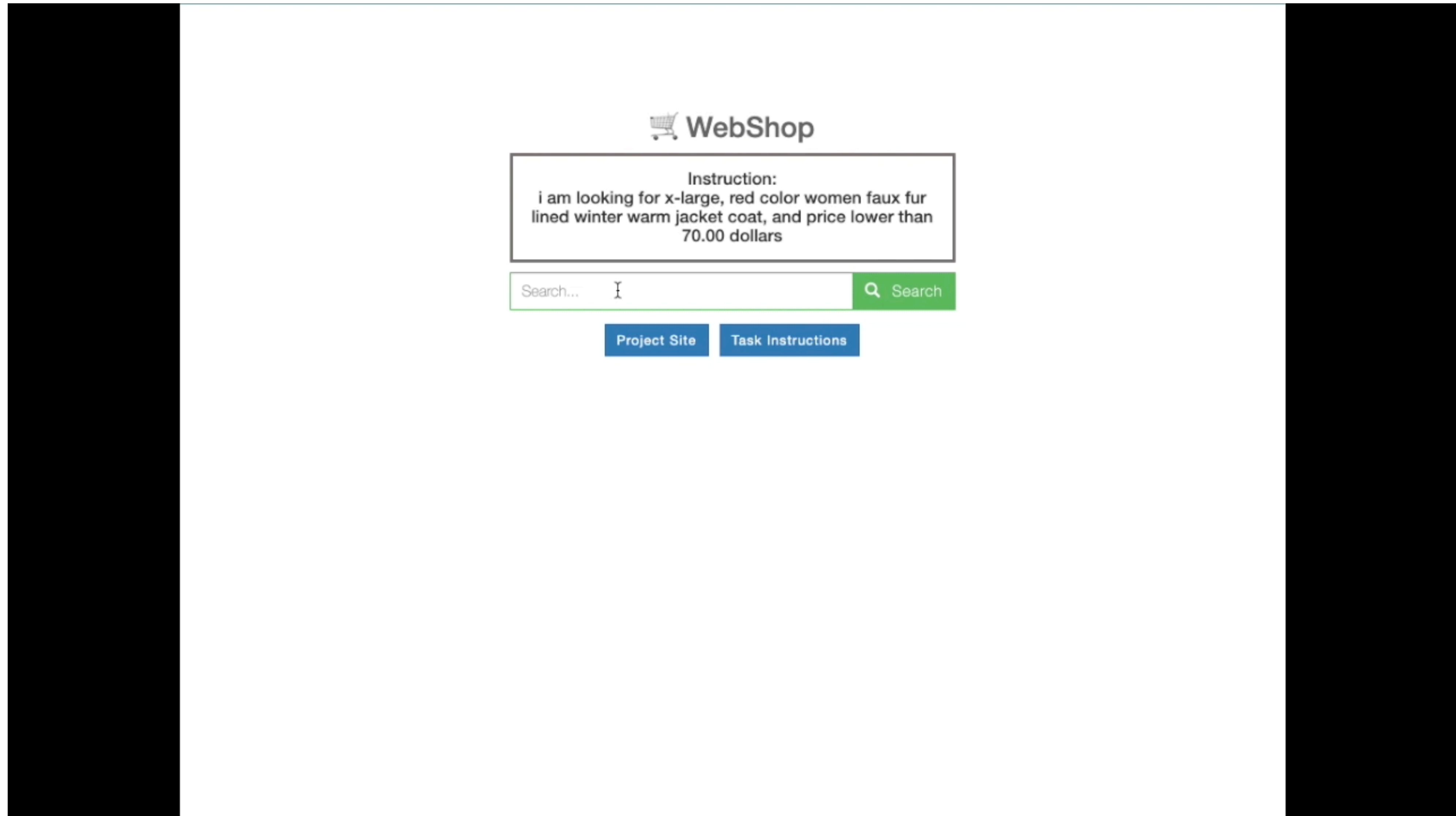
WebShop

A simulated shopping website environment for interactive agents

- **Task:** Given an instruction in natural language, navigate through webpages to purchase the item with the desired requirements

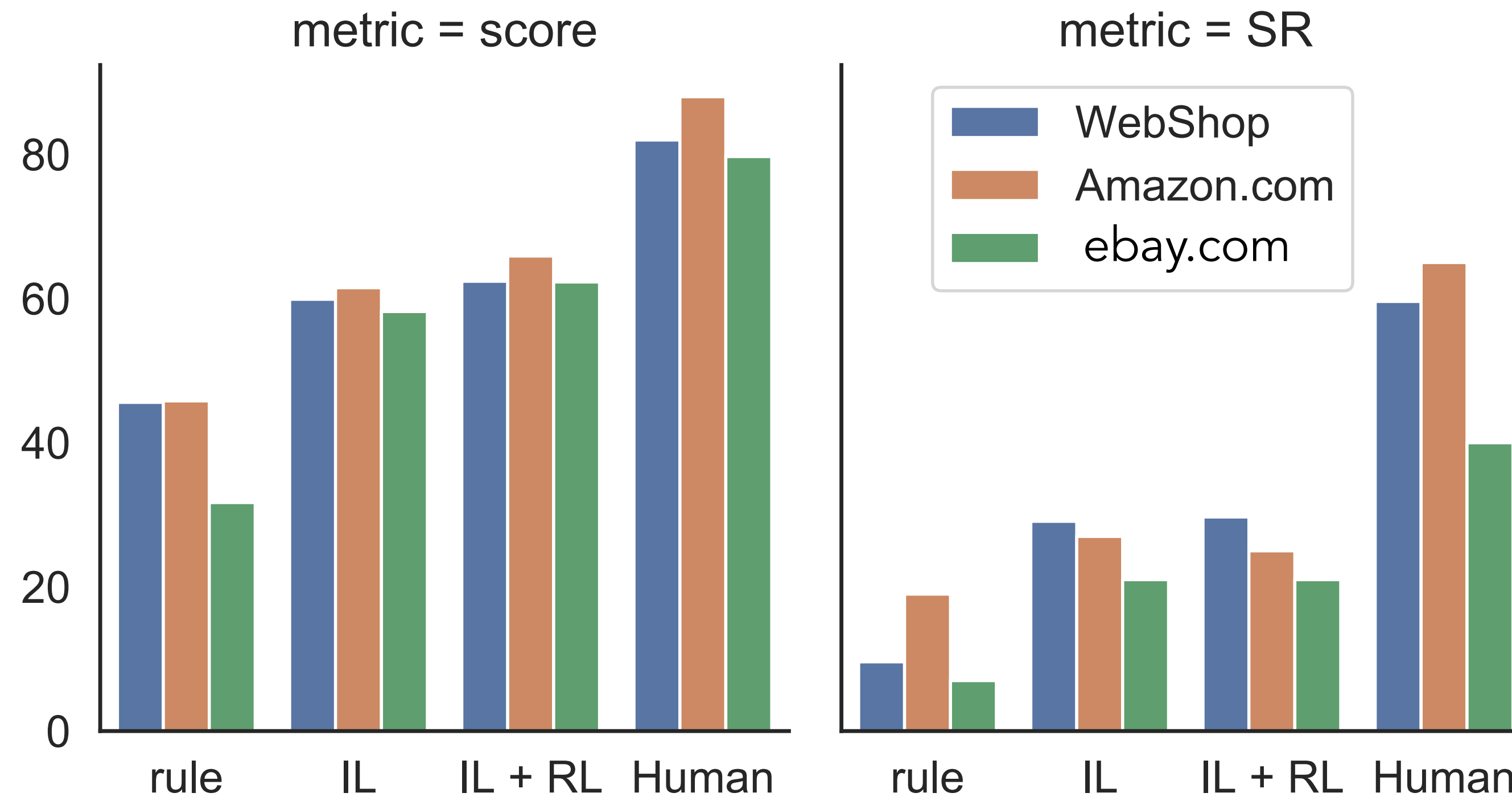


WebShop in a browser

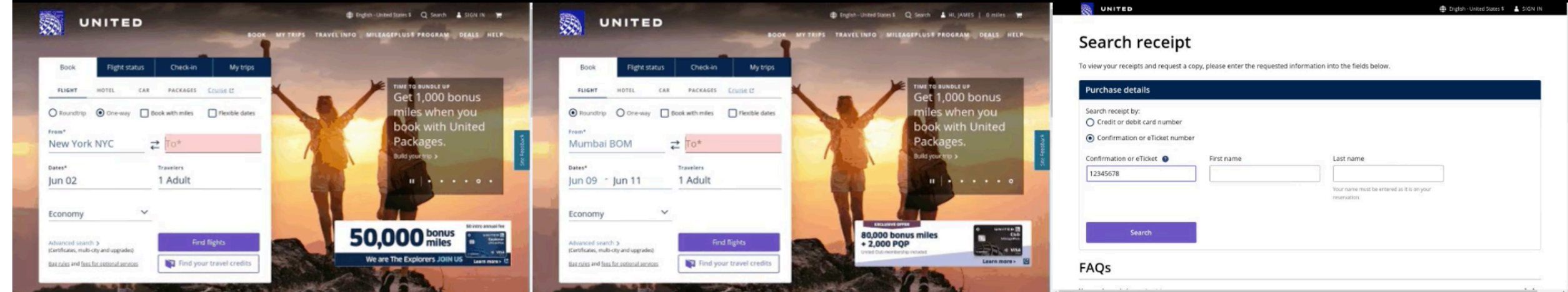


Sim-to-real Transfer

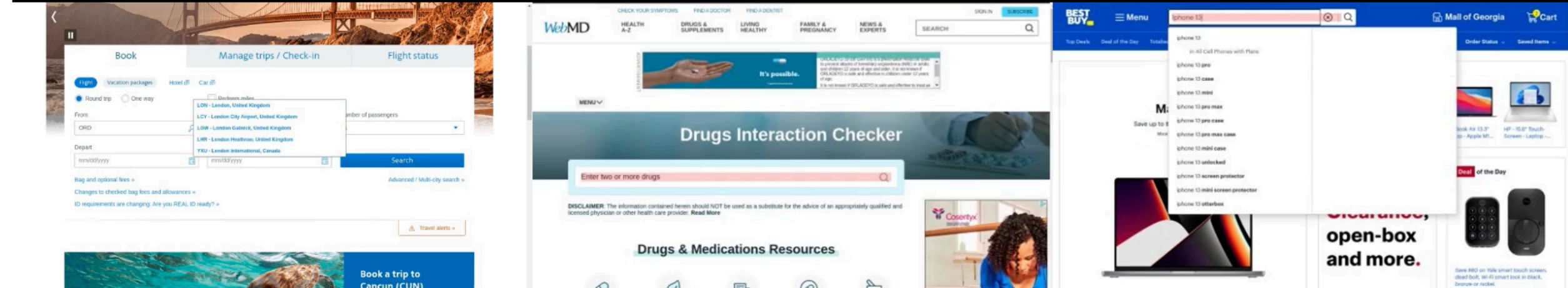
- Deploy models trained on WebShop directly to [amazon.com](https://www.amazon.com) and [ebay.com](https://www.ebay.com)
- Promising transfer to real websites despite differences in products and search engines



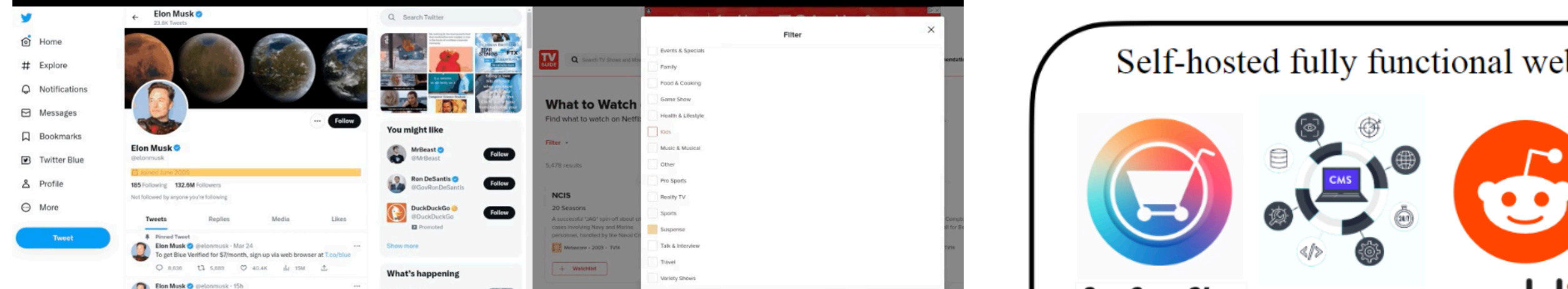
(a) Find one-way flights from New York to Toronto. (b) Book a roundtrip on July 1 from Mumbai to London and vice versa on July 5 for two adults... (c) Search receipt with the eTicket 12345678 for the trip reserved by Jason Two



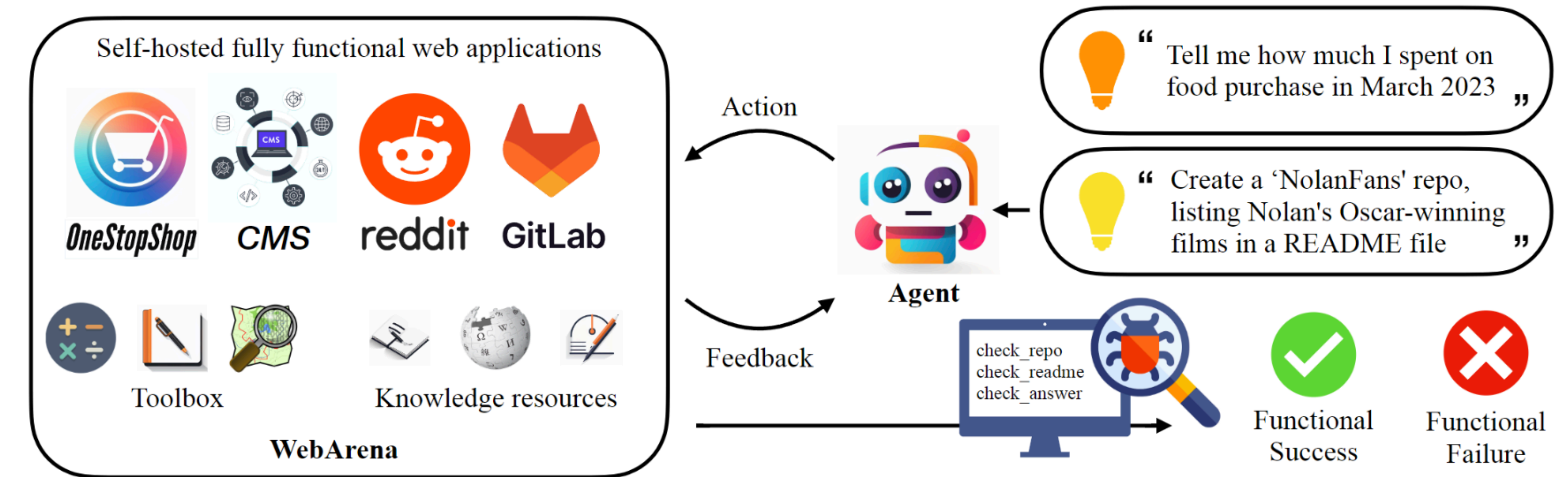
(d) Find a flight from Chicago to London on 20 April and return on 23 April. (e) Search for the interactions between ibuprofen and aspirin. (f) As a Verizon user, finance a blue iPhone 13 with 256gb along with monthly apple care.



(g) Find Elon Musk's profile and start following, start notifications and like the latest tweet. (h) Browse comedy films streaming on Netflix that was released from 1992 to 2007. (i) Open page to schedule an appointment for



Mind2Web (Deng et al., 2023)



WebArena (Zhou et al., 2023)

InterCode: an interactive coding test

Lightweight, flexible, and easy-to-use framework for designing **interactive coding tasks** to **evaluate language agents that can code**.



Standard RL formulation (POMDP), where an **action** is **code**, and **observation** is execution feedback.

InterCode-SQL

Task Setting

Setting: MySQL Database

Dataset: Spider

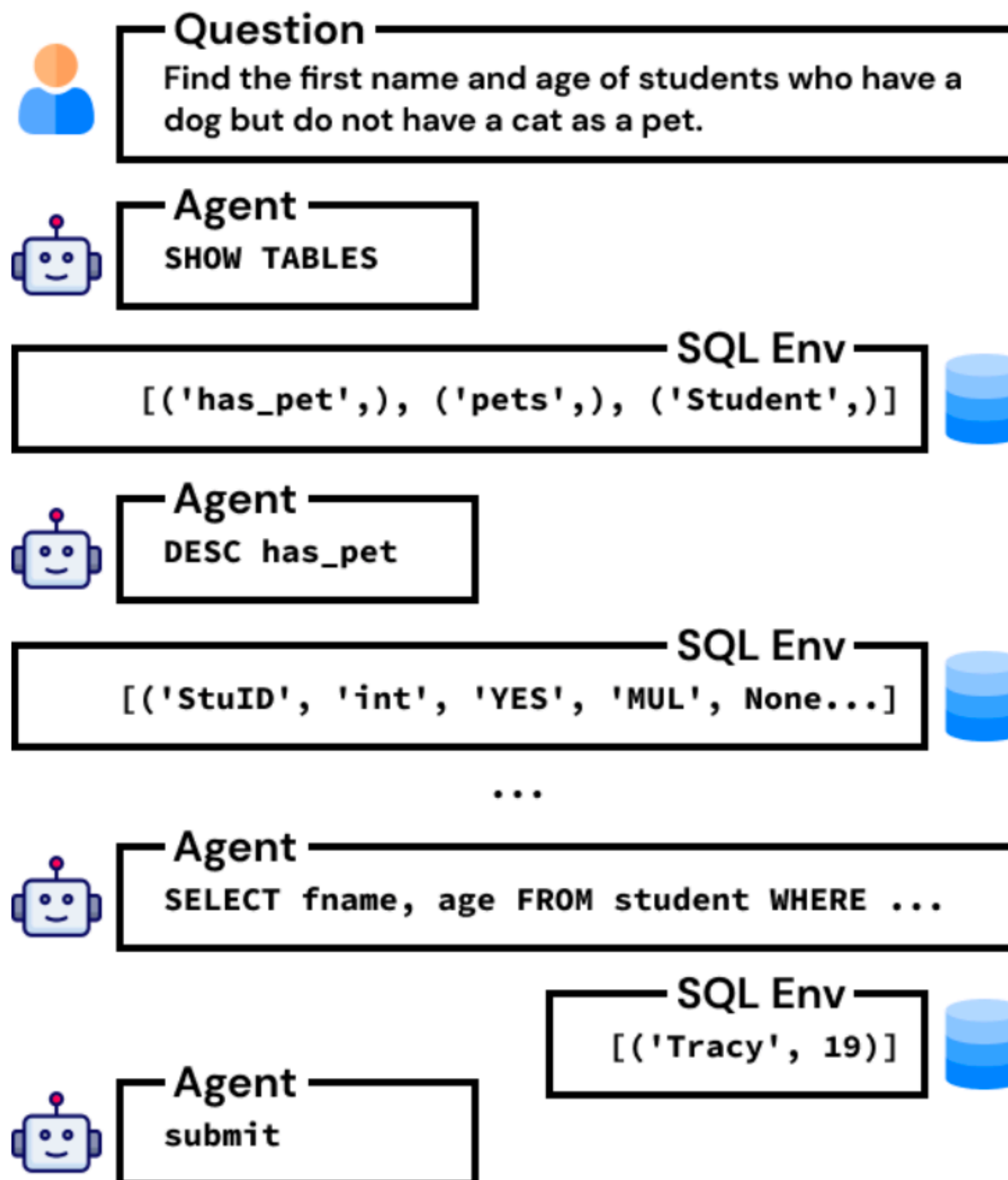
Action Space: SQL commands

Example Task Instance:

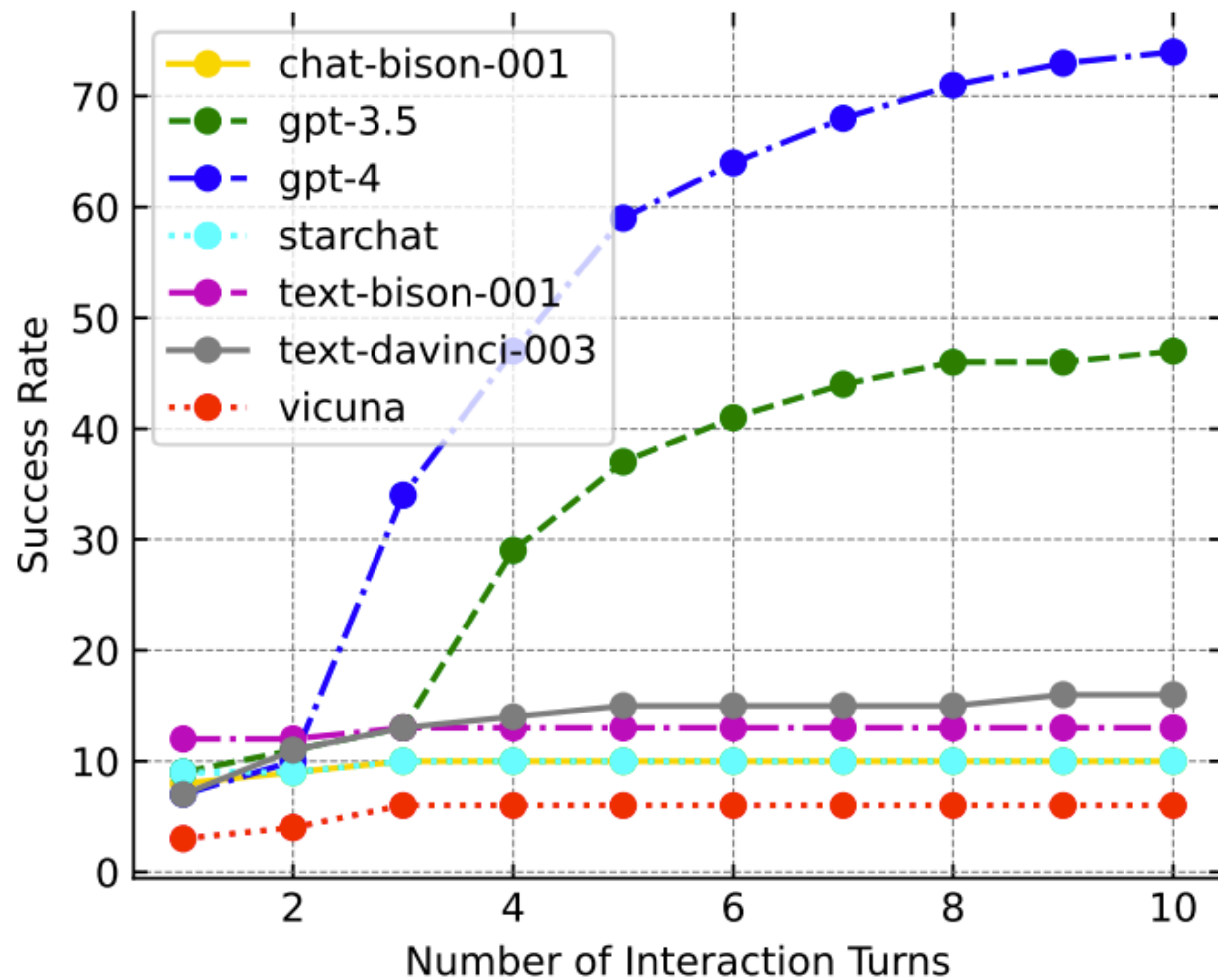
Instruction: Find the first name and age of students who have a dog but not a cat.

Gold: SELECT T1.fname, T1.age FROM student AS T1 JOIN has_pet AS T2 ON T1.stuid = T2.stuid JOIN pets AS T3 ON...

(5 more lines)

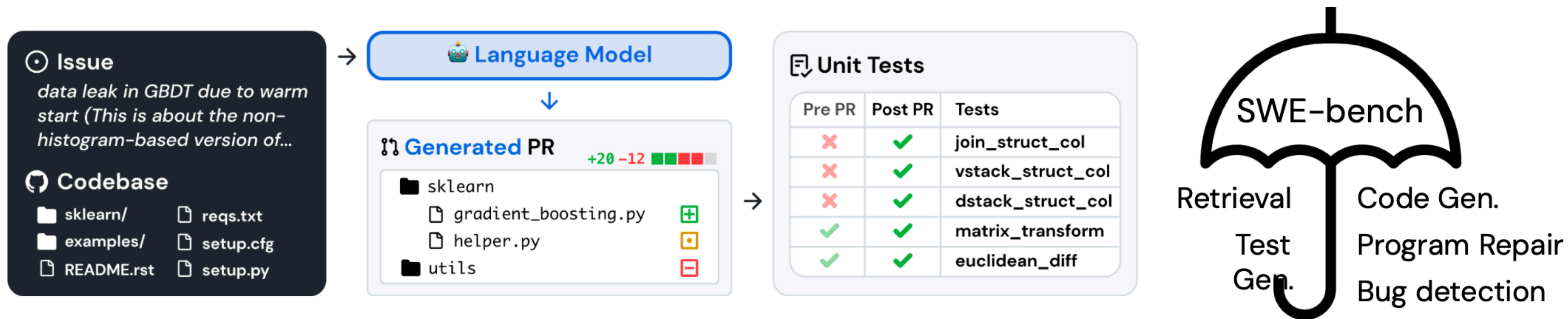


InterCode-SQL



SWE-Bench: a software engineering benchmark

? Given an issue and a codebase, generate a pull request that solves the described issue.



🔧 Use InterCode to construct a terminal + code editor environment.

(Jimenez*, Yang* et al., ICLR, 2024)

Model Input

▼ Instructions

• 1 line

You will be provided with a partial code base and an issue statement explaining a problem to resolve.

▼ Issue

• 67 lines

napoleon_use_param should also affect "other parameters" section Subject: napoleon_use_param should also affect "other parameters" section

Problem

Currently, napoleon always renders the Other parameters section as if napoleon_use_param was False, see source

```
def _parse_other_parameters_section(self, section: str) -> List[str]:
    # type: (unicode) -> List[unicode]
    return self._format_fields(_('Other Parameters'), self._consume_fields())

def _parse_parameters_section(self, section):
    # type: (unicode) -> List[unicode]
    fields = self._consume_fields()
    if self._config.napoleon_use_param: ...
```

▼ Code

• 1431 lines

► README.rst

• 132 lines

► sphinx/ext/napoleon/docstring.py

• 1295 lines

► Additional Instructions

• 57 lines

Gold Patch

sphinx/ext/napoleon/docstring.py

```
def _parse_other_parameters_section(self, section: str) -> List[str]:
- return self._format_fields(_('Other Parameters'), self._consume_fields())
+ if self._config.napoleon_use_param:
+     # Allow to declare multiple parameters at once (ex: x, y: int)
+     fields = self._consume_fields(multiple=True)
+     return self._format_docutils_params(fields)
+ else:
+     fields = self._consume_fields()
+     return self._format_fields(_('Other Parameters'), fields)
```

Generated Patch

sphinx/ext/napoleon/docstring.py

```
def _parse_other_parameters_section(self, section: str) -> List[str]:
- return self._format_fields(_('Other Parameters'), self._consume_fields())
+ return self._format_docutils_params(self._consume_fields())
```

Generated Patch Test Results

```
PASSED NumpyDocstringTest (test_yield_types)
PASSED TestNumpyDocstring (test_escape_args_and_kwargs 1)
PASSED TestNumpyDocstring (test_escape_args_and_kwargs 2)
PASSED TestNumpyDocstring (test_escape_args_and_kwargs 3)
PASSED TestNumpyDocstring (test_pep526_annotations)
FAILED NumpyDocstringTest (test_parameters_with_class_reference)
FAILED TestNumpyDocstring (test_token_type_invalid)
===== 2 failed, 45 passed, 8 warnings in 5.16s =====
```

Model evaluations

Model	BM25 Retrieval		“Oracle” Retrieval	
	% Resolved	% Apply	% Resolved	% Apply
Claude 2	1.96	29.86	4.80	46.99
ChatGPT-3.5	0.20	10.50	0.52	12.38
GPT-4*	0.00	4.50	1.74	13.24
SWE-Llama 7b	0.70	37.84	3.01	54.80
SWE-Llama 13b	0.70	39.41	3.97	52.14

	Claude 2	ChatGPT-3.5	GPT-4*	SWE-Llama 7b	SWE-Llama 13b
Before 2023	4.87	0.49	2.36	2.95	3.98
From 2023	4.23	0.77	0.0	3.46	3.85

Inherently difficult for LLMs to solve; well suited for agents to provide valuable delta

Key challenges

- 1) Evaluations
 - a) Different from how we've built benchmarks for NLP, RL
 - b) Mixture of fields
- 2) Principled frameworks for agent development**
- 3) Mechanisms for ensuring safety and trustworthiness
 - a) Agents much more powerful and dynamic than static NLP models

Production Systems

- Set of rules specifying a precondition + an action
- When a precondition is met, corresponding action can be taken
- Example: String rewriting
 - $XYZ \rightarrow XWZ$
- Can also express entire algorithms (e.g. division by 5)
 - $*|llll \rightarrow |*$
 - $* \rightarrow *$ (terminal)
 - $\rightarrow *$ (zero)

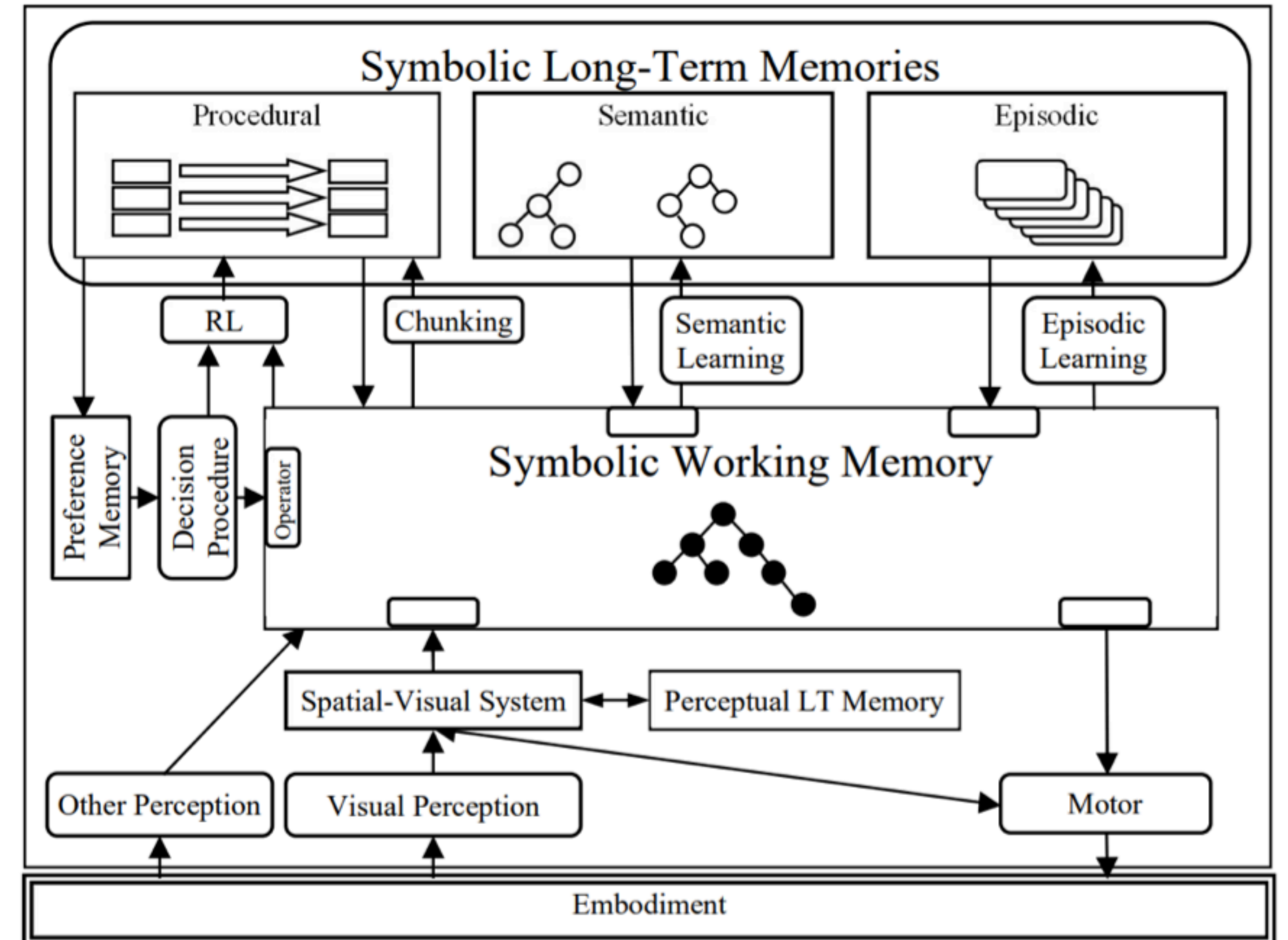
Productions can define agents

- Thermostat controller

$(\text{temperature} > 70^\circ) \wedge (\text{temperature} < 72^\circ) \rightarrow \text{stop}$
 $\text{temperature} < 32^\circ \rightarrow \text{call for repairs; turn on electric heater}$
 $(\text{temperature} < 70^\circ) \wedge (\text{furnace off}) \rightarrow \text{turn on furnace}$
 $(\text{temperature} > 72^\circ) \wedge (\text{furnace on}) \rightarrow \text{turn off furnace}$

Cognitive Architectures build on productions

- Augment productions with:
 - Memory
 - Grounding
 - Decision making procedures
 - Learning mechanisms

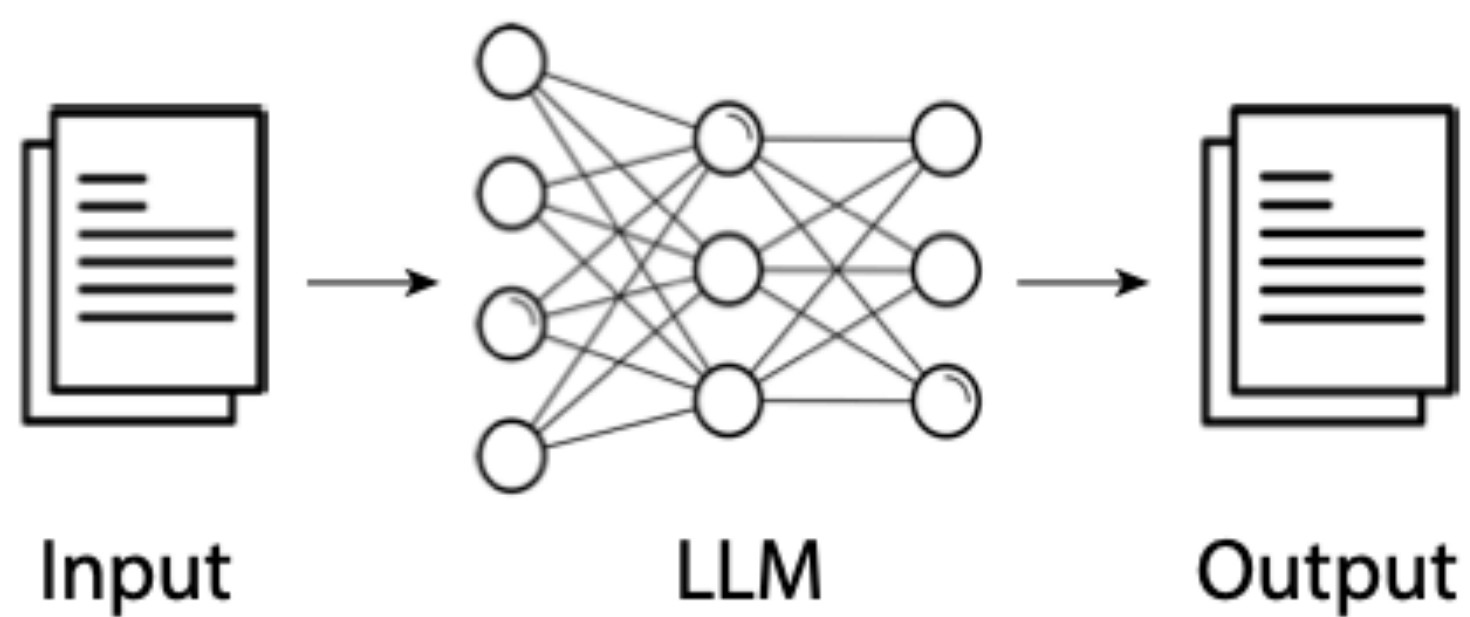
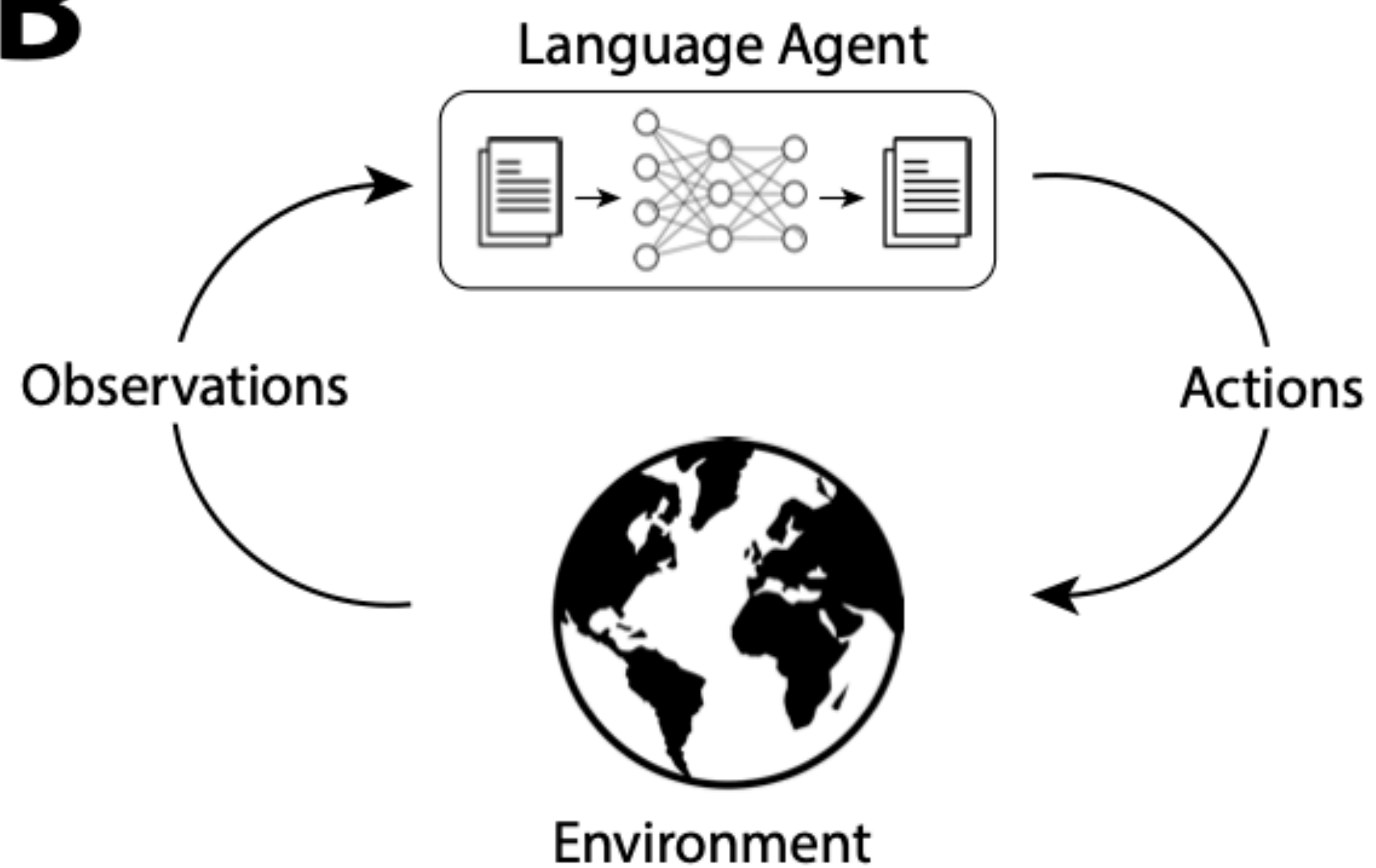
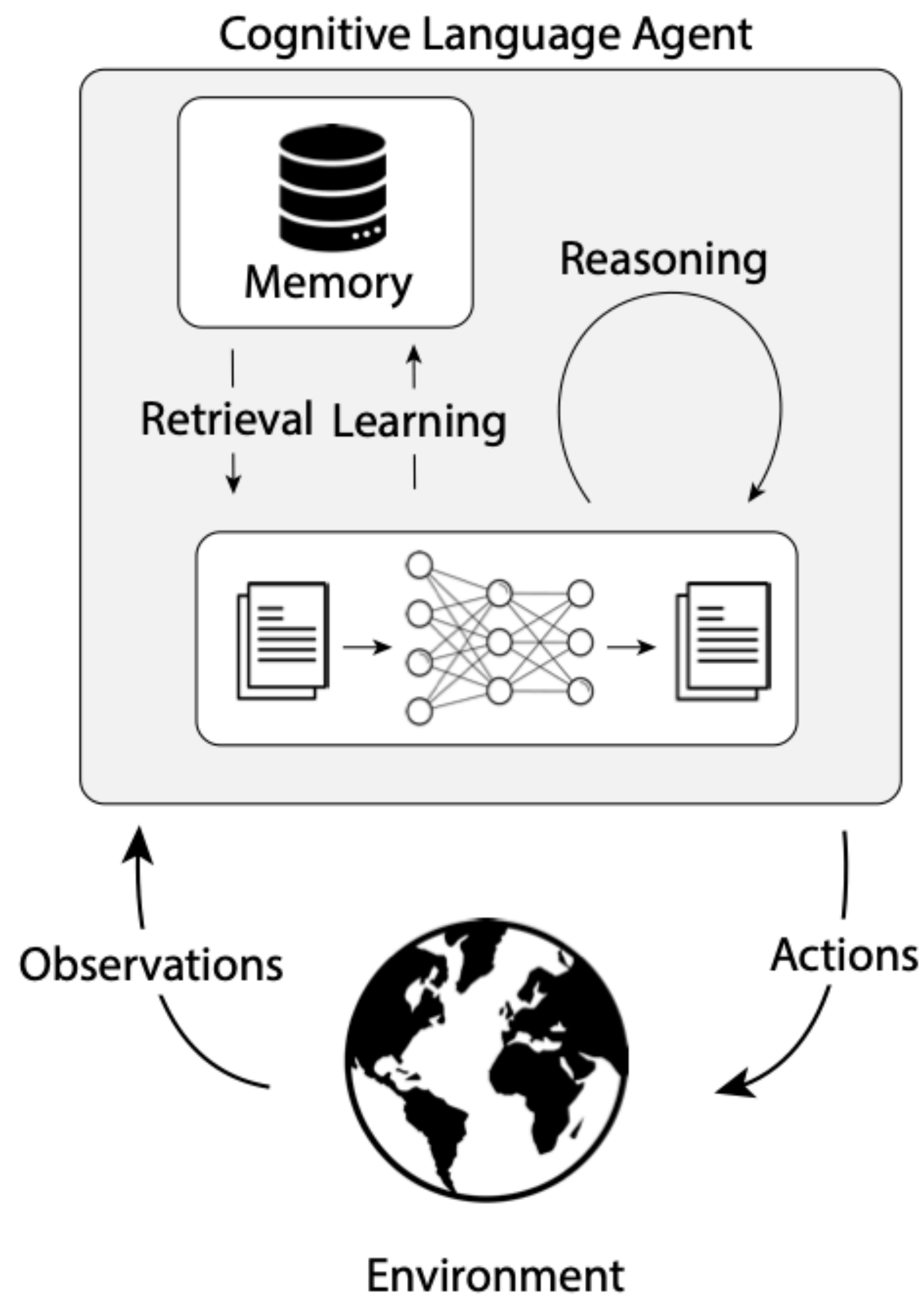


SOAR architecture (Laird et al. 1987, Laird, 2022)

Language Models as Production Systems

- Generate language given a context
 - The cat sat on the \rightarrow mat
- Can be viewed as *probabilistic* production systems
 - $X \rightarrow X Y_i$ according to some distribution $P(Y_i | X)$
 - Represent this as $X \rightsquigarrow X Y$

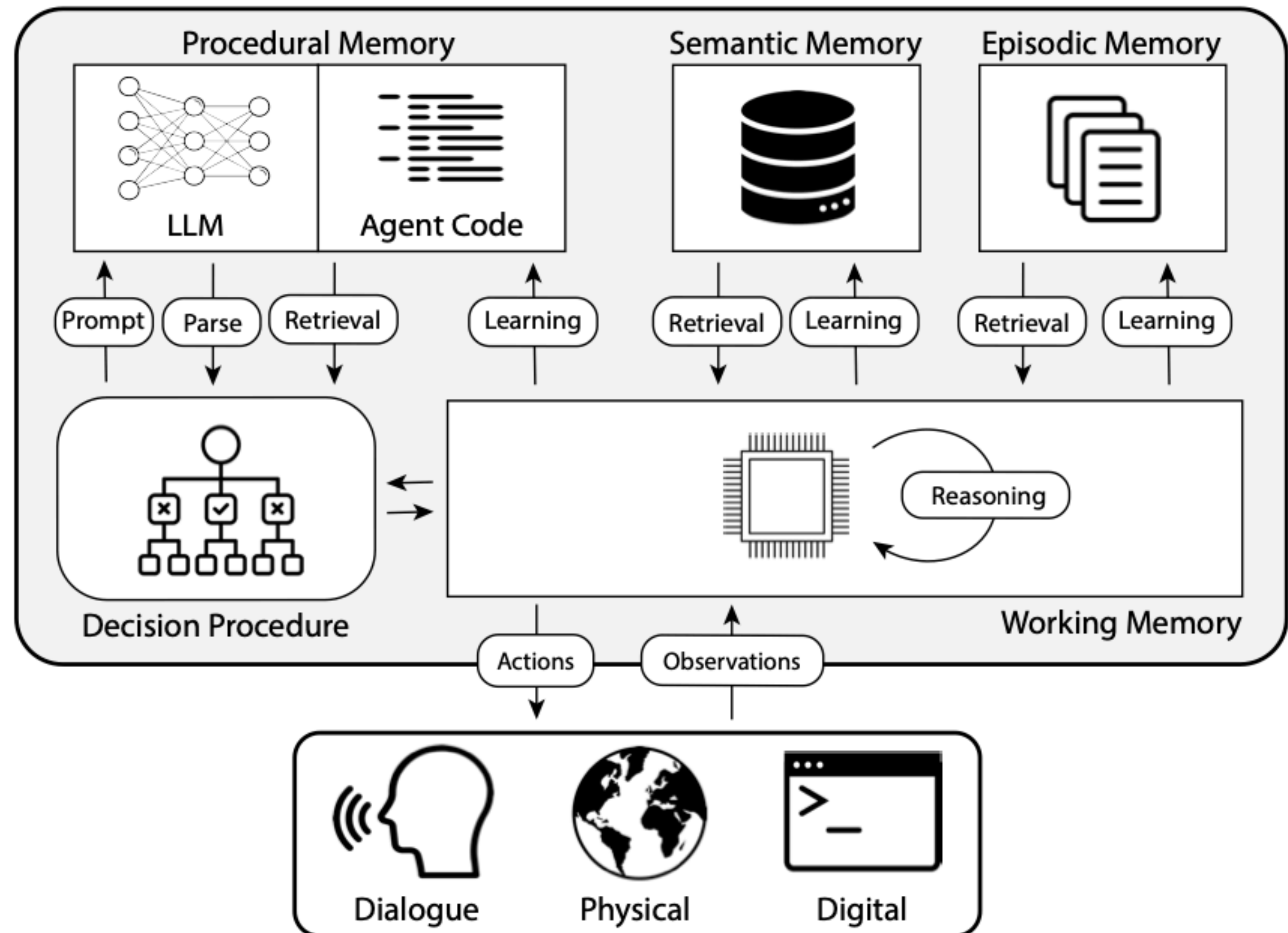
Prompting Method	Production Sequence
Zero-shot	$Q \xrightarrow{\text{LLM}} Q A$
Few-shot (Brown et al., 2020)	$Q \rightarrow Q_1 A_1 Q_2 A_2 Q \xrightarrow{\text{LLM}} Q_1 A_1 Q_2 A_2 Q A$
Zero-shot Chain-of-Thought (Kojima et al., 2022)	$Q \rightarrow Q_{\text{Step-by-step}} \xrightarrow{\text{LLM}} Q_{\text{Step-by-step}} A$
Retrieval Augmented Generation (Lewis et al., 2020)	$Q \xrightarrow{\text{Wiki}} Q O \xrightarrow{\text{LLM}} Q O A$
Socratic Models (Zeng et al., 2022)	$Q \xrightarrow{\text{VLM}} Q O \xrightarrow{\text{LLM}} Q O A$
Self-Critique (Saunders et al., 2022)	$Q \xrightarrow{\text{LLM}} Q A \xrightarrow{\text{LLM}} Q A C \xrightarrow{\text{LLM}} Q A C A$

A**B****C**

CoALA: Cognitive Architectures for Language Agents

- Provide a blueprint for building robust agents
- Categorize existing systems into a single framework
- Key modules: memory, grounding, decision making routines, learning

(Sumers*, Yao* et al., TMLR 2024)



With great(er) power, comes
great(er) responsibility

Technical risks

X-SCITECH >

Microsoft shuts down AI chatbot after it turned into a Nazi

ChatGPT hallucinates court cases

COMPUTING

Racial Bias Found in a Major Health Care Risk Algorithm

- Agents automate entire chain of decisions
- Any single misstep can have repercussions down the pipeline
- Issues like *toxicity, bias* have greater impact
- Hallucination issues can lead to deleted emails, emptied bank accounts, etc.

Societal Risks



- Broader questions around the safe, controlled deployment of AI agents
 - Robustness
 - Malicious use
 - Job displacement
 - Existential risk
- How do we 'align' AI models and agents with human values?

Language Agents in the Digital World: Opportunities and Risks, Yao and Narasimhan (2023)

<https://princeton-nlp.github.io/language-agent-impact>

Summary

- Language agents are the next evolution of language models
- We are closer than ever to the initial dream of *holistic* artificial intelligence
 - Lines between traditional fields like NLP, RL, CV, cognitive science will become blurrier
- Developing robust, useful agents requires good evaluations, principled frameworks and a clear understanding of the risks involved
 - Lots of room for combining ideas from classical AI with modern ML/DL

