



From Language Models to Language Agents

从语言模型到语言智能体

Shunyu Yao

Language Model



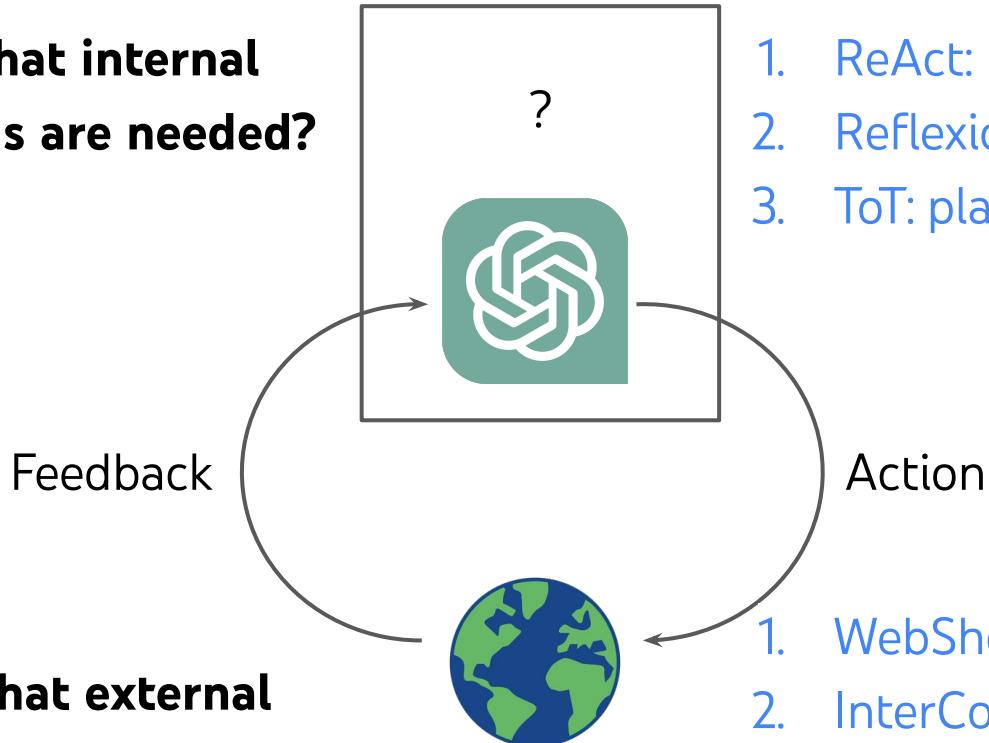
Great, but

- Stateless
- Ungrounded
- Limited Knowledge

Language Agent

Part I: what internal mechanisms are needed?

1. ReAct: reasoning
2. Reflexion: learning
3. ToT: planning



Part II: what external environments are needed?

1. WebShop: web
2. InterCode: code
3. Collie: logic

Part I: Internal Mechanisms

Mechanism 1: Reasoning

Published as a conference paper at ICLR 2023

REACT: SYNERGIZING REASONING AND ACTING IN LANGUAGE MODELS

Shunyu Yao^{*,1}, Jeffrey Zhao², Dian Yu², Nan Du², Izhak Shafran², Karthik Narasimhan¹, Yuan Cao²

¹Department of Computer Science, Princeton University

²Google Research, Brain team

Reasoning helps Acting

Obs t

You are cooking a dish and seeing salt is out...

Reasoning

"The dish should be savory, and since salt is out, I should find the soy sauce instead. It is in the cabinet to my right..."

Act t

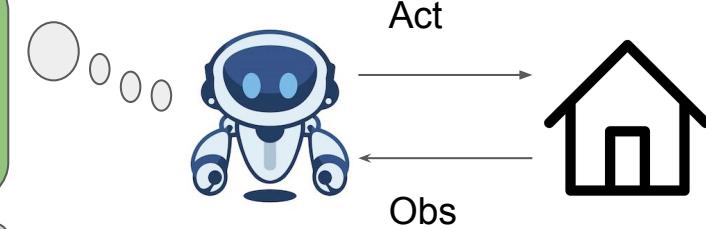
Turn right

Obs t+1

You see a cabinet and a table...

Act t+1

Open cabinet



Reason how to act

1. Generalization
2. Alignment

.....
“Verbal Reasoning” as Augmented Action Space!

Acting helps Reasoning

SH

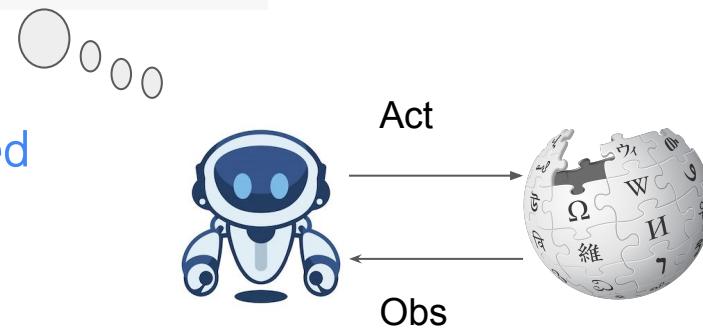
who is PM of UK?



As of December 2022, the Prime Minister of the United Kingdom is Boris Johnson.

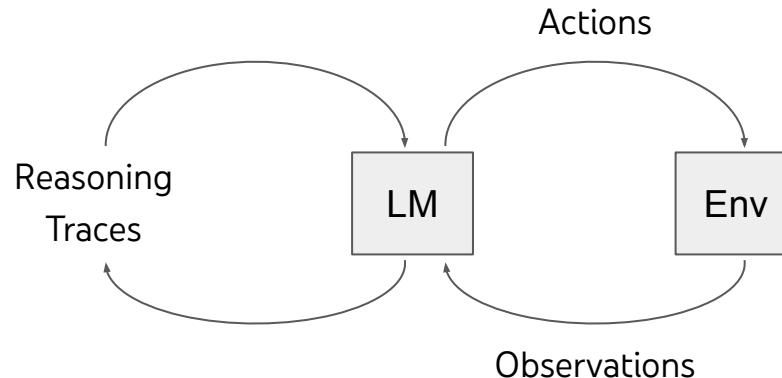
However, since I am a large language model trained by OpenAI, my knowledge is limited to what was available at the time of my training in 2021. My ability to provide current information is limited, and I am not able to browse the internet for up-to-date information.

Internal knowledge can be limited / outdated
Hallucination is serious
Acting is key



Reasoning and Acting

- Language models (LM) get better at reasoning and acting, **separately**
- **ReAct**: Let LMs generate both
 - **Reasoning traces** (to update **internal** model belief)
 - **Actions** (to obtain **external** environment feedback)



ReAct (Reason + Act)

ReAct: Overview

- **Tasks:** Question answering, Fact verification, Text game, Web Interaction
- **Learning:** prompting / finetuning
- **Model:** PaLM-540B / GPT-3

- **Synergy:** Reasoning guides acting, acting supports reasoning
- **Benefits of ReAct**
 - Flexibility: diverse reasoning / interactive tasks
 - Generalization: strong few-shot / fine-tuning performances
 - Alignment: the **human** way of problem solving!

ReAct Prompting and Controlled Baselines

ReAct prompt

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]

Reason-only prompt
(Chain-of-thought, CoT)

Question Which magazine was started first Arthur's Magazine or First for Women?
Thought Arthur's Magazine was started in 1844. First for Women was started in 1989. 1844 (Arthur's Magazine) < 1989 (First for Women), so Arthur's Magazine was started first.
Answer Arthur's Magazine

Act-only prompt
(similar to WebGPT)

Question Which magazine was started first Arthur's Magazine or First for Women?
Action 1 Search[Arthur's Magazine]
Observation 1 Arthur's Magazine (1844-1846) was an American literary periodical published in Philadelphia ...
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...
Action 3 Finish[Arthur's Magazine]

Standard prompt

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

Without Act: Misinformation

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

Standard

Answer: iPod

X

Reason Only

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

X

ReAct: Interpretable, Factual

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

ReAct (Reason + Act)

Thought 1: I need to search Apple Remote and find the program it was originally designed to interact with.

Act 1: `Search[Apple Remote]`

Obs 1: The Apple Remote is a remote control introduced in October 2005 by Apple ... originally designed to control the `Front Row media center` program ...

Thought 2: Apple Remote was originally designed to control the Front Row media center program. I need to search `Front Row` next and find what other device can control it.

Act 2: `Search[Front Row]`

Obs 2: Could not find `[Front Row]`. Similar: `['Front Row Seat to Earth', 'Front Row Motorsports', 'Front Row (software)', ...]`

Thought 3: Front Row is not found. I need to search `Front Row (software)`.

Act 3: `Search[Front Row (software)]`

Obs 3: Front Row is a discontinued media center software ...

Thought 4: `Front Row (software)` is controlled by an Apple Remote or the keyboard function keys So the answer is keyboard function keys.

Act 4: `Finish[keyboard function keys]`



Act Only: Unable to Synthesize Final Answer

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

Act Only

Act 1: `Search[Apple Remote]`

Obs 1: The Apple Remote is a remote control introduced in October 2005 by Apple ... originally designed to control the `Front Row media center` program ...

Act 2: `Search[Front Row]`

Obs 2: Could not find [Front Row]. Similar: ['Front Row Seat to Earth', 'Front Row Motorsports', 'Front Row (software)', ...]

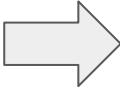
Act 3: `Search[Front Row (software)]`

Obs 3: Front Row is a discontinued media center software ...

Act 4: `Finish[yes]`

X

Embodied Tasks: ALFWorld



Embodied Env
(ALFRED game)

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.

> take vase 2 from shelf 6

You pick up the vase 2 from the shelf 6.

> go to safe 1

You arrive at loc 3. The safe 1 is closed.

> open safe 1

You open the safe 1. The safe 1 is open. In it, you see a keychain 3.

> put vase 2 in/on safe 1

You won!

Text Env
(ALFWorld)

ALFWorld Example: Reasoning is Key to Long-horizon Acting

You are in the middle of a room. Looking quickly around you, you see a cabinet 6, a cabinet 1, a coffee machine 1, a countertop 3, a stove burner 1, and a toaster 1.

Your task is to: Put some pepper shaker on a drawer.

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 dishponge 2, and a spoon 1.

Act 3: Go to sinkbasin 1

Obs 3: On the sinkbasin 1, you see a dishponge 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.



ALFWorld Example: Reasoning is Key to Long-horizon Acting

You are in the middle of a room. Looking quickly around you, you see a cabinet 6, a cabinet 1, a coffee machine 1, a countertop 3, a stove burner 1, and a toaster 1.

Your task is to: Put some pepper shaker on a drawer.

ReAct (Reason + Act)

Act 1: Think[First I need to find a pepper shaker...more likely to appear in cabinets (1-6), countertops (1-3), ...]

Act 2: Go to cabinet 1

Obs 2: On the cabinet 1, you see a vase 2.

(Here the agent goes to cabinet 1, 2, 3, 4, 5, 6, then countertop 1 and 2)

Act 9: Go to countertop 3

Obs 9: On the countertop 3, you see a apple 1, a bread 1, a pepper shaker 1, and a vase 3.

Act 10: Take pepper shaker 1 from countertop 3

Obs 10: You pick up the pepper shaker 1 from the countertop 3.

Act 11: Think[Now I find a pepper shaker 1. Next, I need to put it in/on drawer 1].

Act 12: Go to drawer 1

Obs 12: Drawer 1 is closed.

Act 13: Open drawer 1

Obs 13: You open Drawer 1 ...

Act 14: Put pepper shaker 1 in/on drawer 1

Obs 14: You put pepper shaker 1 in/on the drawer 1.



ALFWorld Example: Human-in-the-loop Control

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.



ALFWorld Example: Human-in-the-loop Control

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.

(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.

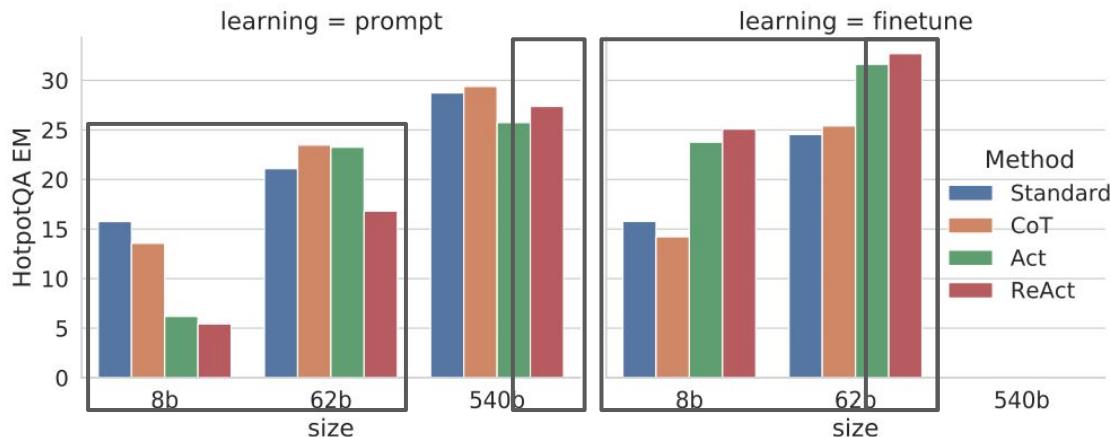
Act 29: put keychain 2 in/on safe 1

Obs 29: You put the keychain 2 in/on the safe 1.



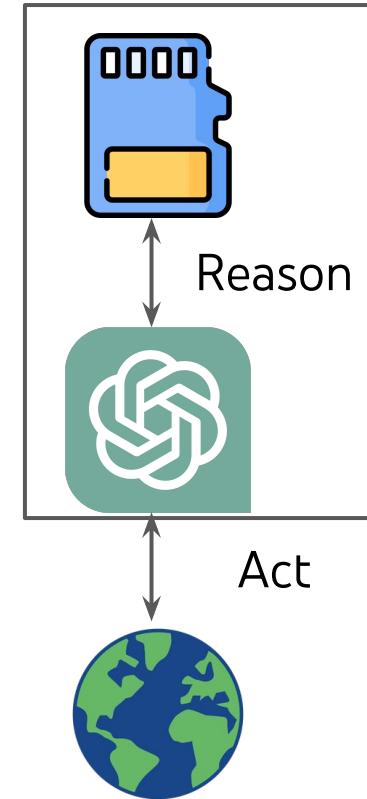
Finetuning > Prompting

- Prompting only works with LLMs, limited learning support
- Finetuning is promising (initial results on HotpotQA, using prompted trajectories)
 - ReAct finetuned small LMs > ReAct prompted large LMs
 - ReAct finetuning is better than other finetuning formats **across model sizes**

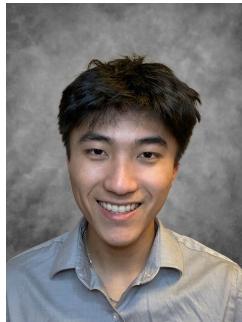


ReAct: Summary

- Two basic functions of language agents: reason and act
 - Many followups since 2022: LangChain, AutoGPT, ...
- Simple view: **action space**
- Slightly more complex view: **working memory**
 - LM can “talk” to memory (reason) or env (act)
 - Memory short-term, append-only
 - What if memory is long-term, not append-only?



Mechanism 2: Learning



Reflexion: Language Agents with Verbal Reinforcement Learning

Noah Shinn

Northeastern University
noahshinn024@gmail.com

Federico Cassano

Northeastern University
cassano.f@northeastern.edu

Slides credit

Beck Labash

Northeastern University
labash.b@northeastern.edu

Ashwin Gopinath

Massachusetts Institute of Technology
agopi@mit.edu

Karthik Narasimhan

Princeton University
karthikn@princeton.edu

Shunyu Yao

Princeton University
shunyuy@princeton.edu

(Reinforcement) Learning

- Behavior -> Feedback -> Update -> Better Behavior

Learning: Feedback

- Behavior -> **Feedback** -> Update -> Better Behavior

```
for i in people.data.users:
    response = client.api.statuses.user_timeline.get(screen_name=i.screen_name)
    print 'Got', len(response.data), 'tweets from', i.screen_name
    if len(response.data) != 0:
        ldate = response.data[0]['created_at']
        ldate2 = datetime.strptime(ldate, '%a %b %d %H:%M:%S +0000 %Y')
        today = datetime.now()
        howlong = (today-ldate2).days
        if howlong < daywindow:
            print i.screen_name, 'has tweeted in the past', daywindow,
            tottweets += len(response.data)
            for j in response.data:
                if j.entities.urls:
                    for k in j.entities.urls:
                        newurl = k['expanded_url']
                        urlset.add((newurl, j.user.screen_name))
        else:
            print i.screen_name, 'has not tweeted in the past', daywindo
```

- **Scalar Feedback:** 0 or 1
- **Language Feedback:**
 - Runtime error messages
 - Unit test cases and results
 - Human/self-feedback: “the code did not handle this corner case...”

Learning: Update

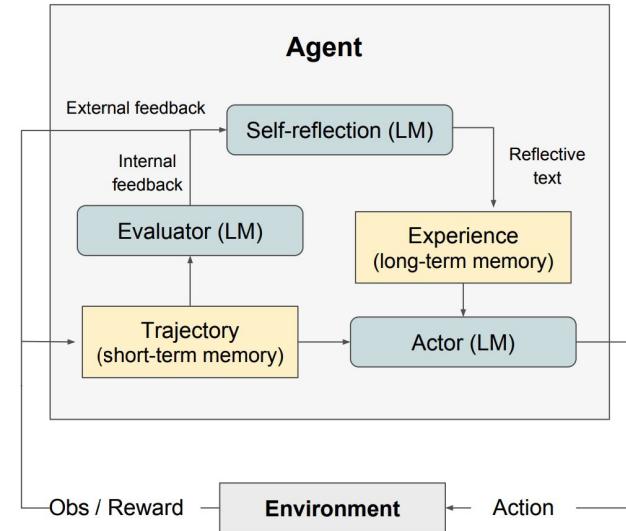
- Behavior -> Feedback -> **Update** -> Better Behavior

```
for i in people.data.users:  
    response = client.api.statuses.user_timeline.get(screen_name=i.screen_name)  
    print 'Got', len(response.data), 'tweets from', i.screen_name  
    if len(response.data) != 0:  
        ldate = response.data[0]['created_at']  
        ldate2 = datetime.strptime(ldate, '%a %b %d %H:%M:%S +0000 %Y')  
        today = datetime.now()  
        howlong = (today-ldate2).days  
        if howlong < daywindow:  
            print i.screen_name, 'has tweeted in the past' , daywindow,  
            tottweets += len(response.data)  
            for j in response.data:  
                if j.entities.urls:  
                    for k in j.entities.urls:  
                        newurl = k['expanded_url']  
                        urlset.add((newurl, j.user.screen_name))  
        else:  
            print i.screen_name, 'has not tweeted in the past', daywind
```

- **Parameter Update:** PPO, A3C, DQN ...
- **Language Update:**
 - “Be sure to handle this corner case...”

Reflexion: “Verbal” RL

- RL: Behavior -> Feedback -> Update -> Better Behavior
 - Traditional RL: scalar feedback, parameter update
 - “Verbal” RL: language feedback, language update
- **Key idea:** LM can self-evaluate/reflect/critic
 - Works with various tasks (game, QA, coding, ...)
 - Works with various models (ReAct, CoT, ...)
 - Works with various feedback sources (env, human, self-generated, ...)
 - **Significantly improves the agent across episodes, using long-term memory update**



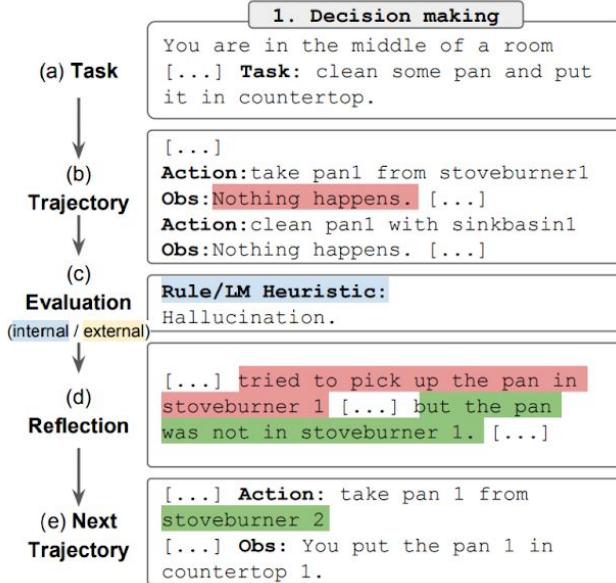
SH

please reflect and see if your solution is correct

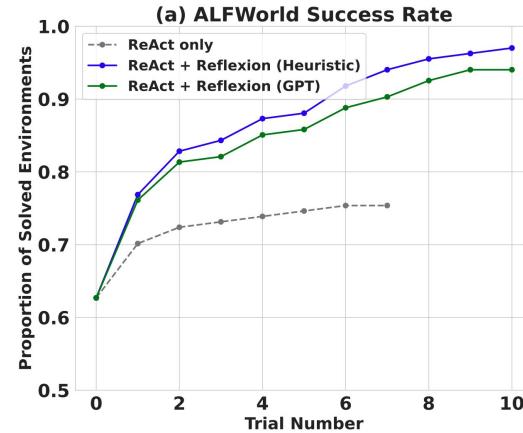


I apologize for any confusion in my previous response. Upon reflection, recursive solution I provided is indeed not optimal for larger positions,

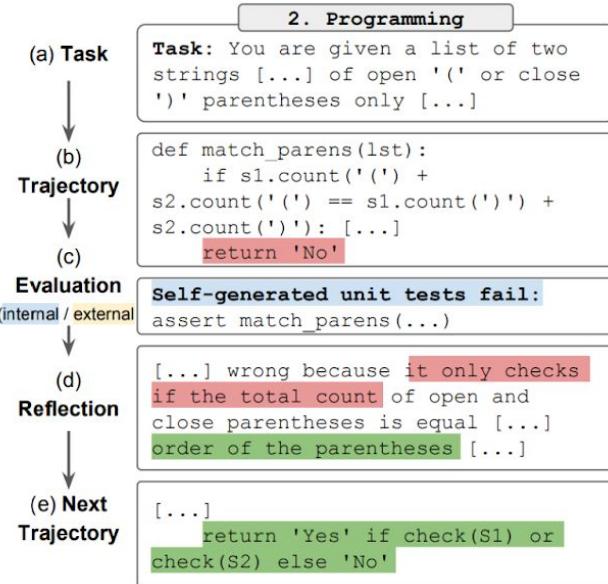
Decision Making (ALFWorld)



1. (Task)
2. (Action + observation trajectory)
3. (Evaluation)
4. (Self-reflection)
5. (Repeat)



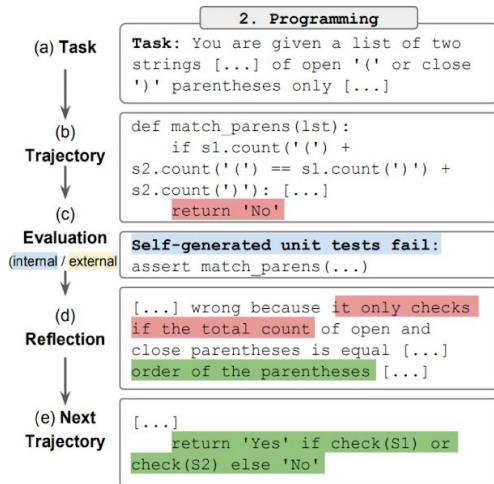
Coding (multi-task, multi-lingual)



Benchmark + Language	Prev SOTA Pass@1	SOTA Pass@1	Reflexion Pass@1
HumanEval (PY)	65.8 (CodeT [4] + GPT-3.5)	80.1 (GPT-4)	91.0
HumanEval (RS)	–	60.0 (GPT-4)	68.0
MBPP (PY)	67.7 (CodeT [4] + Codex [5])	80.1 (GPT-4)	77.1
MBPP (RS)	–	70.9 (GPT-4)	75.4
Leetcode Hard (PY)	–	7.5 (GPT-4)	15.0

1. (Instruction)
2. (Function implementation)
3. (Unit test feedback) ← execution
4. (Self-reflection)
5. (Repeat)

Importance of Unit Test



Test generation omission

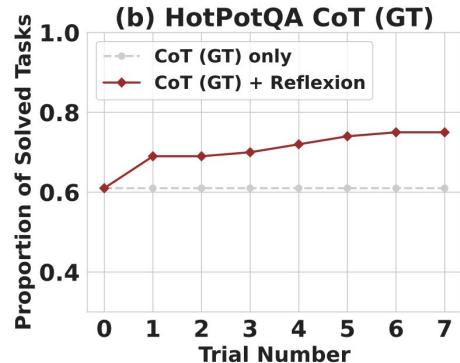
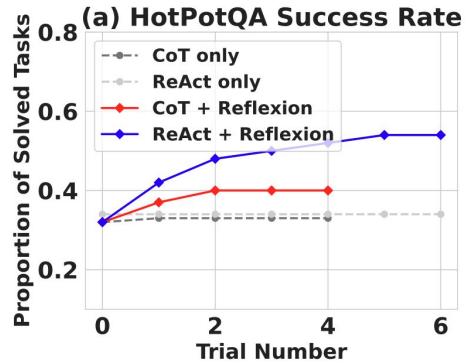
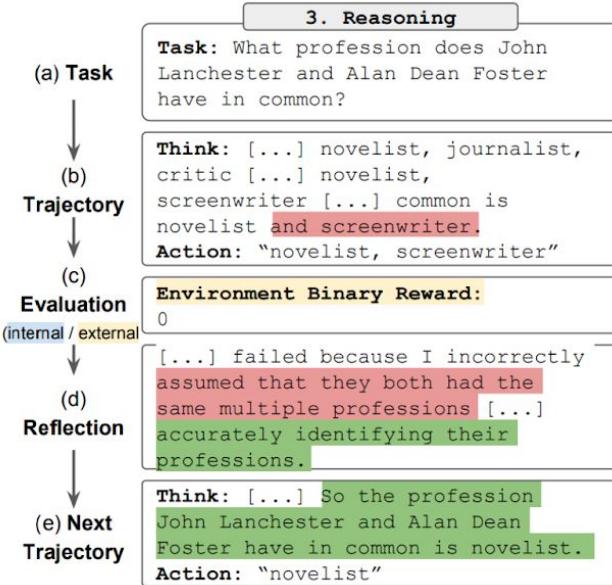
1. (Instruction)
2. (Function implementation)
3. ~~(Unit test feedback) ← execution~~
4. (Self-reflection)
5. (Repeat)

Self-reflection omission

1. (Instruction)
2. (Function implementation)
3. (Unit test feedback) ← execution
4. ~~(Self-reflection)~~
5. (Repeat)

Approach	Test Generation	Self-reflection	Pass@1 (Acc)
Base model	False	False	0.60
Test generation omission	False	True	0.52
Self-reflection omission	True	False	0.60
Reflexion	True	True	0.68

Reasoning (HotpotQA)

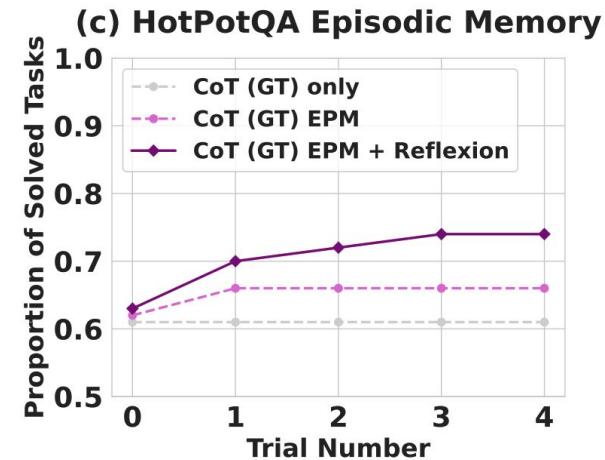
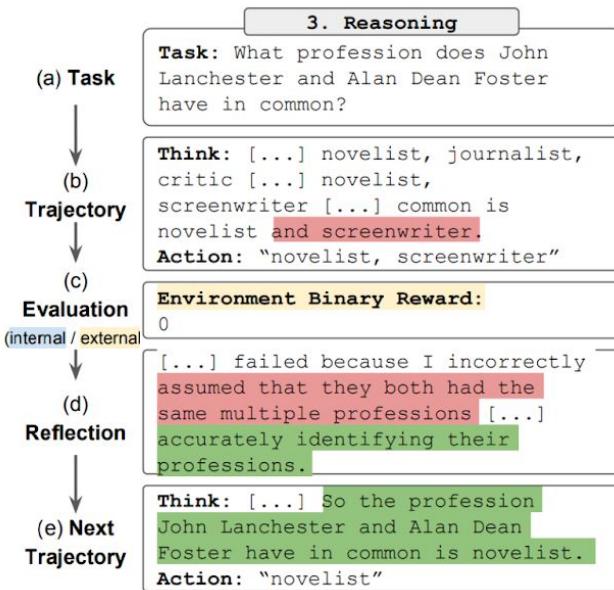


1. (Question)
2. (Knowledge retrieval)
3. (Answer)
4. (Exact match evaluation)
5. (Self-reflection)
6. (Repeat)

Wei, Jason, et al. "Chain of thought prompting elicits reasoning in large language models." arXiv preprint arXiv:2201.11903 (2022).

Yao, Shunyu, et al. "React: Synergizing reasoning and acting in language models." arXiv preprint arXiv:2210.03629 (2022).

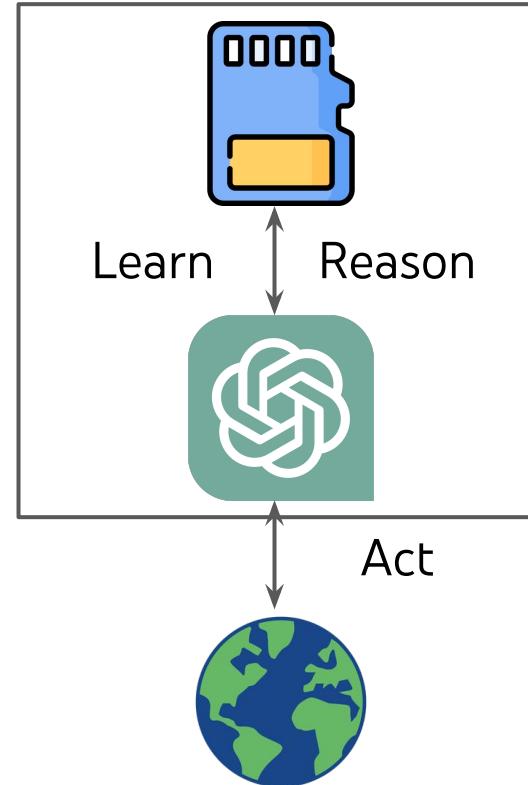
Importance of Reflection



1. (Question)
2. (Knowledge retrieval)
3. (Answer)
4. (Exact match evaluation)
5. (Self-reflection)
6. (Repeat)

Reflexion: Summary

- A verbal / “no-gradient” approach to learning
 - Concurrent work: self-refine, self-debugging, ...
- Simple view: **append reflection to prompt**
- Slightly more complex view: **long-term memory**
- Limitation?



Mechanism 3: Planning

Tree of Thoughts: Deliberate Problem Solving with Large Language Models

Shunyu Yao
Princeton University

Dian Yu
Google DeepMind

Jeffrey Zhao
Google DeepMind

Izhak Shafran
Google DeepMind

Thomas L. Griffiths
Princeton University

Yuan Cao
Google DeepMind

Karthik Narasimhan
Princeton University

Motivating Task: Game of 24

Use numbers and basic arithmetic operations (+ - * /) to obtain 24. Each step, you are only allowed to choose two of the remaining numbers to obtain a new number.

Input: 4 4 6 8

Steps: $4 + 8 = 12$ (left: 4 6 12); $6 - 4 = 2$ (left: 2 12); $2 * 12 = 24$ (left: 24)

Answer: $(6 - 4) * (4 + 8) = 24$

{more CoT examples...}

Input: 4 5 6 10

Steps: $5 - 4 = 1$ (left: 1 6 10); $10 - 1 = 9$ (left: 6 9); $6 * 4 = 24$ (left: 24)

Answer: $(10 - (5 - 4)) * 6 = 24$

Remember: LM was designed to generate text autoregressively, not to act or make decisions!

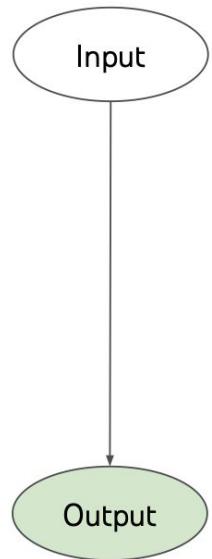
Method	Success
IO prompt	7.3%
CoT prompt	4.0%
CoT-SC ($k=100$)	9.0%
IO + Refine ($k=10$)	27%
ToT (ours) ($b=1$)	45%
ToT (ours) ($b=5$)	74%

Why hard even for GPT-4 + CoT?

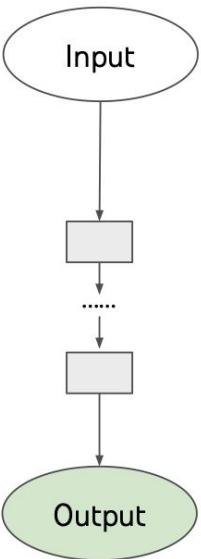
No backtrack: First couple tokens (“5 -”) wrong, game over... (In fact, 60% CoT samples fail after first step)

Token decision: really hard to tell if first token “5” is good or bad

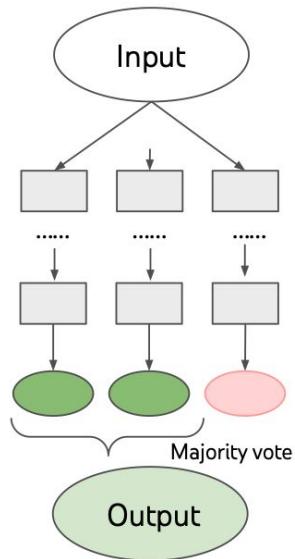
Tree of Thoughts (ToT)



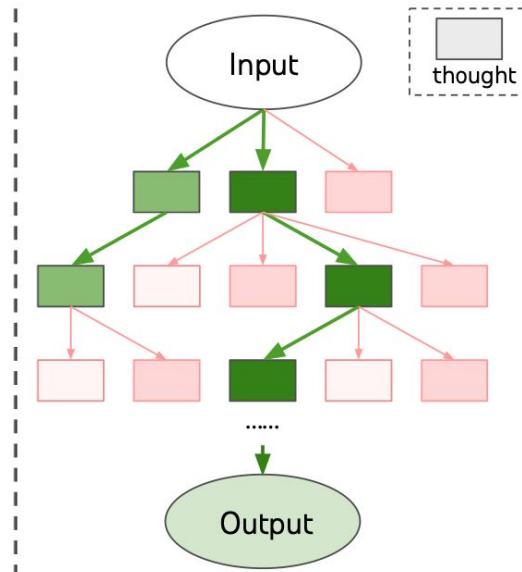
(a) Input-Output
Prompting (IO)



(c) Chain of Thought
Prompting (CoT)



(c) Self Consistency
with CoT (CoT-SC)



(d) Tree of Thoughts (ToT)

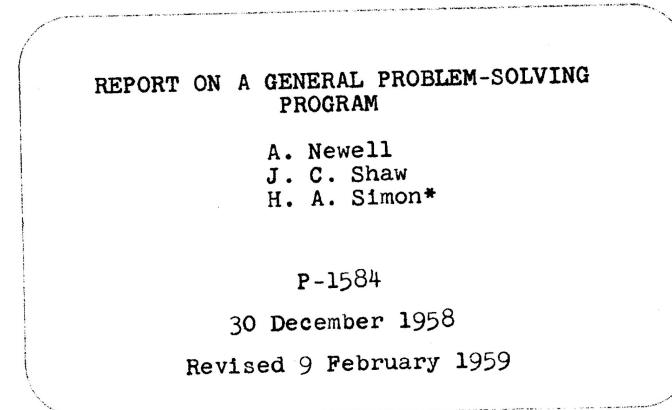
ToT recipe:

1. What is thought?
2. How to generate it?
3. How to evaluate it?
4. How to search?

0. Why Tree Search?

*"A **genuine problem-solving** process involves the repeated use of available information to initiate **exploration**, which discloses, in turn, more information until a way to attain the solution is finally discovered."*

*"Many kinds of information can aid in solving problems: information may suggest the order in which possible solutions should be examined; it may rule out a whole class of solutions ... All these kinds of information are **heuristics** --- things that aid discovery."*

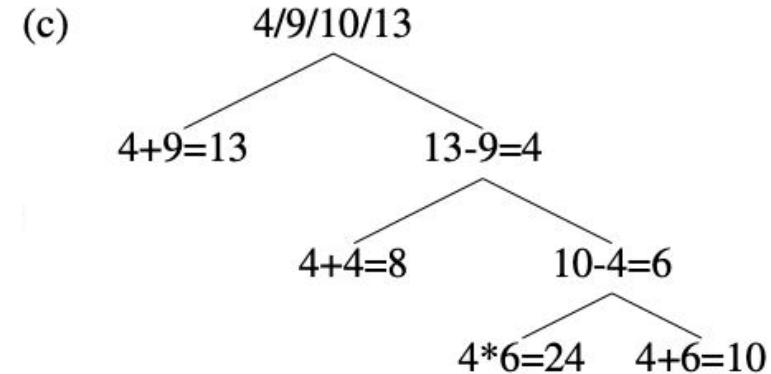
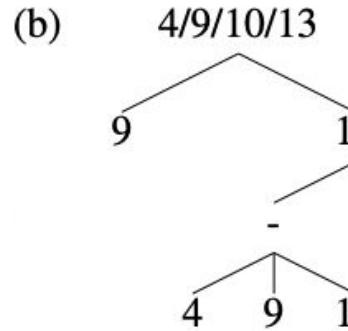


- At the root of AI, tree search was invented as a general problem solving process
- Hand-written heuristics limited the generality of symbolic AI
- Marry the tree search's power as algorithm + LLM's power as heuristics

1. What is thought?

(a) $4/9/10/13$

```
graph TD; A[4/9/10/13] --> B[4+9=13]; A --> C[9-4=5]; B --> D[13-10=3]; B --> E[10/5=2]; C --> F[3*13=39]; C --> G[2*13=26]
```



Bandit of outputs

- Easy to evaluate
- Hard to generate

Tree of tokens

- Easy to generate
- Hard to evaluate

Tree of thoughts

A nice tradeoff game between
generation vs evaluation

2. How to generate thoughts?

Option A. Sample i.i.d. Thoughts

- Good for long thoughts
- Might generate repetitive short thoughts

Use numbers and basic arithmetic operations (+ - * /) to obtain 24. Each step, you are only allowed to choose two of the remaining numbers to obtain a new number.

Input: 4 4 6 8

Steps:

$4 + 8 = 12$ (left: 4 6 12)

$6 - 4 = 2$ (left: 2 12)

$2 * 12 = 24$ (left: 24)

Answer: $(6 - 4) * (4 + 8) = 24$

Input: 2 9 10 12

Steps:

4 runs

GPT-4
(stop='\n')

$4 * 6 = 24$ (left: 5 10 24)

$5 * 4 = 20$ (left: 6 10 20)

$4 * 5 = 20$ (left: 6 10 20)

$5 * 4 = 20$ (left: 6 10 20)

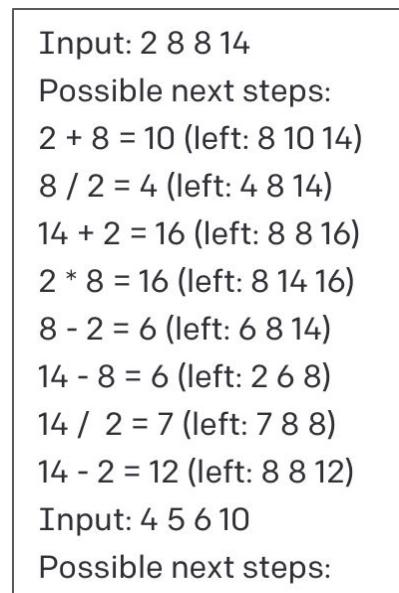
2. How to generate thoughts?

Option A. Sample i.i.d. Thoughts

- Good for long thoughts
- Might generate repetitive short thoughts

Option B. Propose sequential thoughts

- Make sure thoughts are unique in the same context



GPT-4

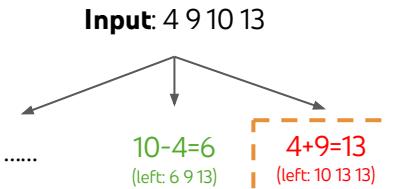
1 run

4 + 5 = 9 (left: 6 9 10)
5 + 6 = 11 (left: 4 10 11)
10 - 4 = 6 (left: 5 6 6)
4 * 5 = 20 (left: 6 10 20)
10 / 5 = 2 (left: 2 4 6)
6 - 4 = 2 (left: 5 2 10)
10 - 6 = 4 (left: 4 4 5)
6 / 4 = 1.5 (left: 5 1.5 10)

*3. How to evaluate thoughts?

Option A. Value independently

- When final “success” is tractable
- More priors injectable
 - E.g. Promote with simulation success
 - E.g. Reject with commonsense



(b) Value Prompt

Evaluate if given numbers can reach 24 (sure/likely/impossible)
10 14: $10 + 14 = 24$. sure
{more examples}
10 13 13

Thought Evaluation

$(13 - 10) * 13 = 3 * 13 = 39$
 $10 + 13 + 13 = 36$
There is no way to obtain 24 with these big numbers. impossible

*3. How to evaluate thoughts?

Option A. Value independently

- When final “success” is tractable
- More priors injectable
 - E.g. Promote with simulation success
 - E.g. Reject with commonsense

Option B. Vote across

- When “success” is open-ended
- Stepwise self-consistency / democracy
- Can be zero-shot

Given an instruction and several choices, decide which choice is most promising. Analyze each choice in detail, then conclude in the last line "The best choice is {s}", where s the integer id of the choice.

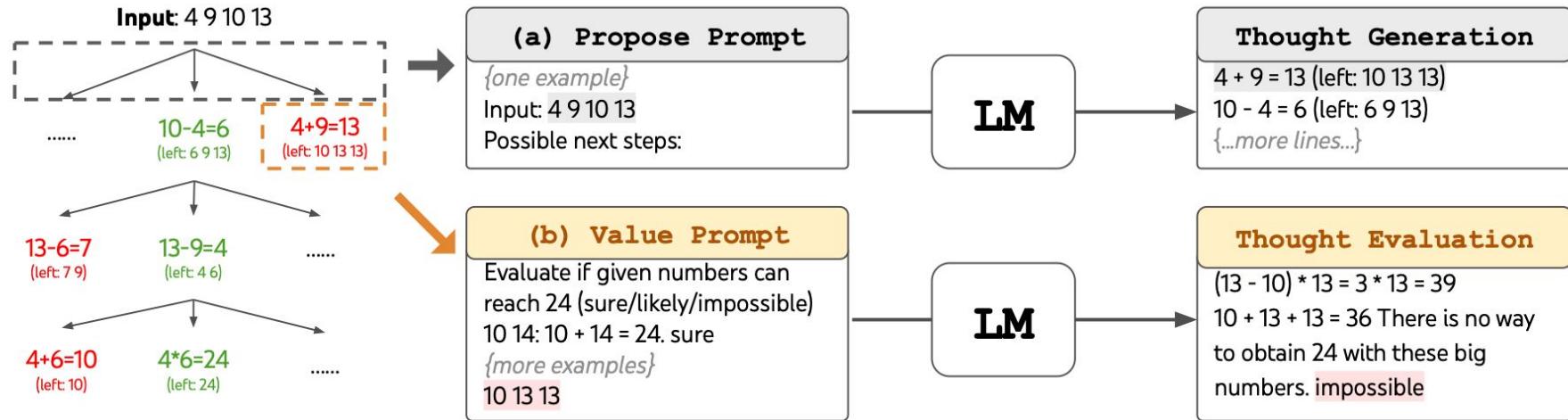
4. How to search?

- Breadth-first search (BFS)
 - Maintain a set of promising thoughts per step
 - Branch and prune per step
 - **Good for:** problems with shallow trees (e.g. ≤ 3 steps),
but hard to get initial choices right
- Depth-first search (DFS)
 - Alway explore the most promising thought
 - Until self-evaluation is “bad”, then backtrack
 - **Good for:** problems with deep trees (e.g. ≥ 10 steps), but
possible to get initial choices right

Tasks

	Game of 24	Creative Writing	5x5 Crosswords
Input	4 numbers (4 9 10 13)	4 random sentences	10 clues (h1. presented;..)
Output	An equation to reach 24 $(13-9)*(10-4)=24$	A passage of 4 paragraphs ending in the 4 sentences	5x5 letters: SHOWN; WIRRA; AVAIL; ...
Thoughts	3 intermediate equations $(13-9=4 \text{ (left } 4,4,10\text{); } 10-4=6 \text{ (left } 4,6\text{); } 4*6=24)$	A short writing plan (1. Introduce a book that connects...)	Words to fill in for clues: (h1. shown; v5. naled; ...)
#ToT steps	3	1	5-10 (variable)

Game of 24



1. Thought: each intermediate equation
2. Generate thoughts: **propose**
3. Evaluate thoughts: **value**
4. Search algorithm: **BFS** (depth=4, breadth<=5)

Game of 24: Results

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.

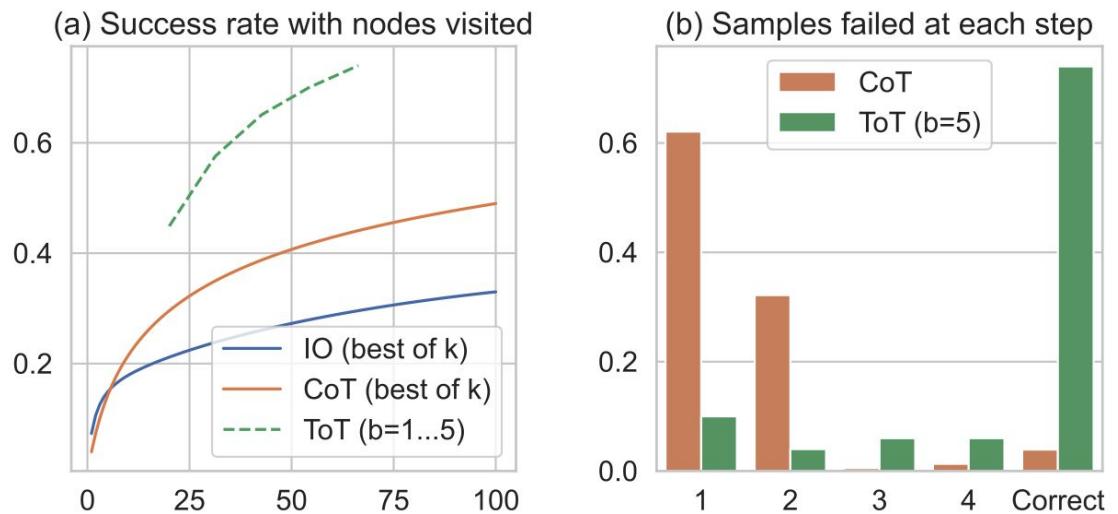


Figure 3: Game of 24 (a) scale analysis & (b) error analysis.

Creative Writing

Write a coherent passage of 4 short paragraphs. The end sentence of each paragraph must be:

- 1.** It isn't difficult to do a handstand if you just stand on your hands.
- 2.** It caught him off guard that space smelled of seared steak.
- 3.** When she didn't like a guy who was trying to pick her up, she started using sign language.
- 4.** Each person who knows you has a different perception of who you are.

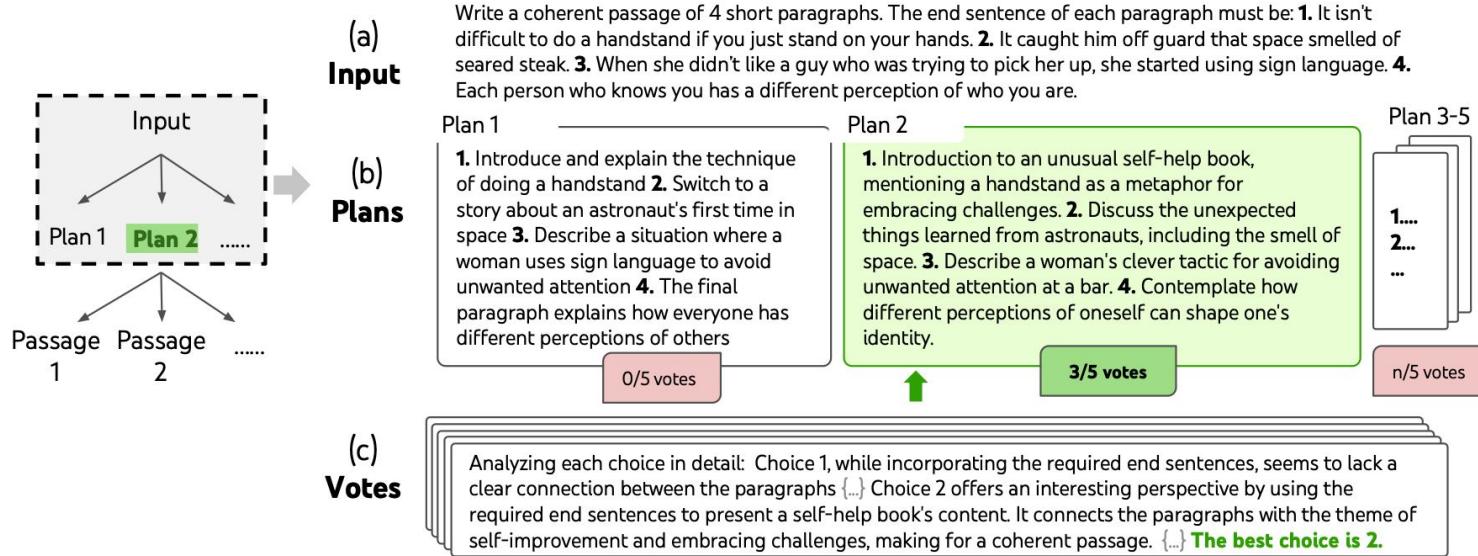
Plan 1

- 1.** Introduce and explain the technique of doing a handstand
- 2.** Switch to a story about an astronaut's first time in space
- 3.** Describe a situation where a woman uses sign language to avoid unwanted attention
- 4.** The final paragraph explains how everyone has different perceptions of others

Plan 2

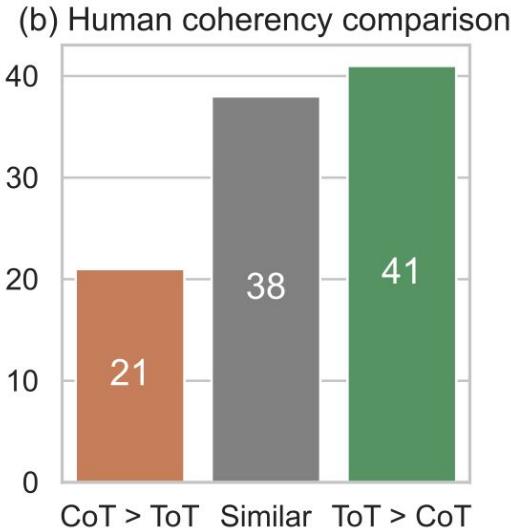
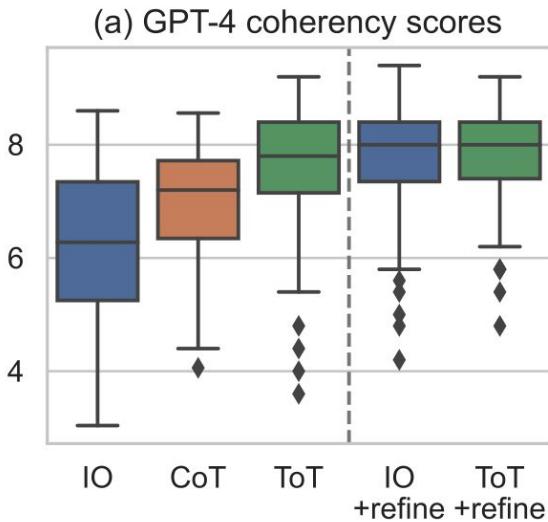
- 1.** Introduction to an unusual self-help book, mentioning a handstand as a metaphor for embracing challenges.
- 2.** Discuss the unexpected things learned from astronauts, including the smell of space.
- 3.** Describe a woman's clever tactic for avoiding unwanted attention at a bar.
- 4.** Contemplate how different perceptions of oneself can shape one's identity.

Creative Writing



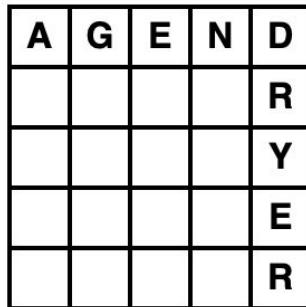
1. Thought: just one intermediate plan
2. Generate thoughts: **sample**
3. Evaluate thoughts: **vote**
4. Search algorithm: **BFS** (depth=2, breadth=5)

Creative Writing: Results



Mini Crosswords

<https://www.goobix.com/crosswords/0505/1/>



In order to indicate an answer, you must click on the number associated with its definition.

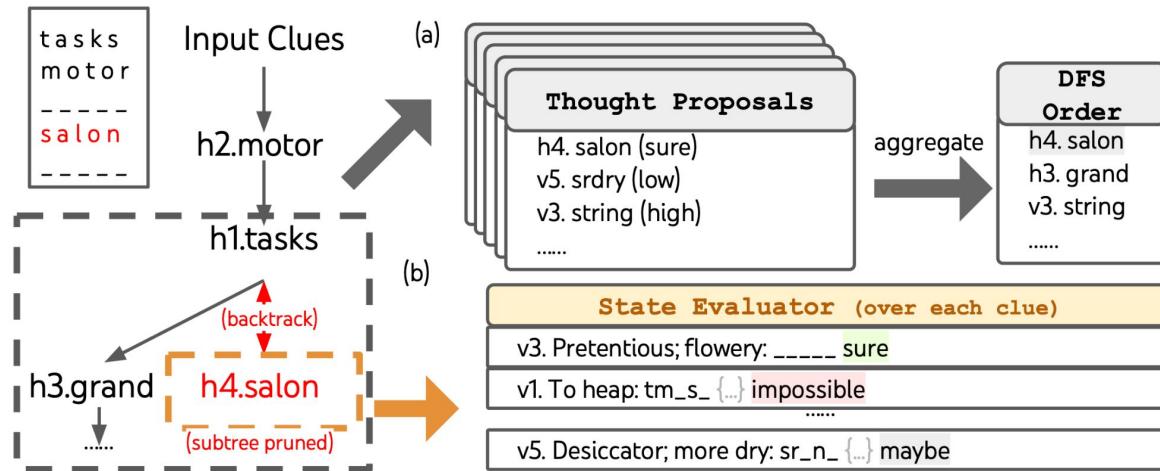
Horizontal

- 1 An agendum; something to be done
- 2 An engine
- 3 Pretentious; flowery
- 4 A salon; a hall
- 5 To mock; to sneer

Vertical

- 1 To heap
- 2 An Indian antelope
- 3 To intend; to plan; to devise; a nettle; to guess
- 4 A nozzle
- 5 Desiccator; more dry

Mini Crosswords



1. Thought: each clue word
2. Generate thoughts: **propose**
3. Evaluate thoughts: **value**
4. Search algorithm: **DFS** (depth=5-10)

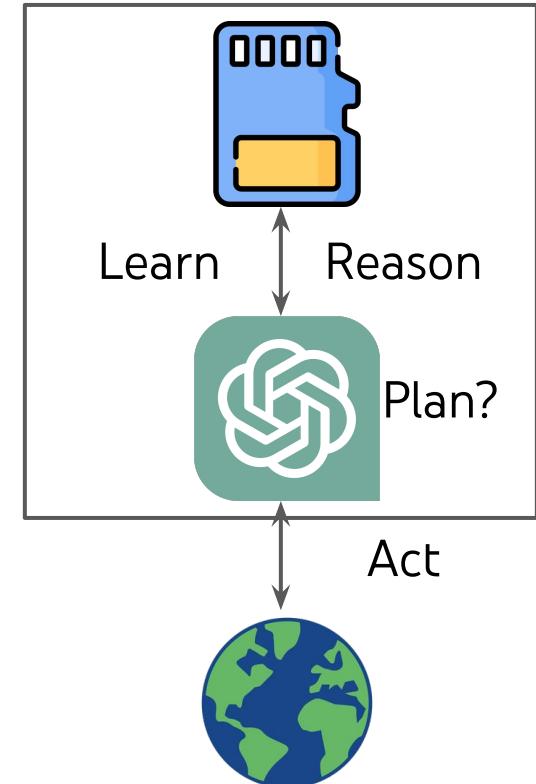
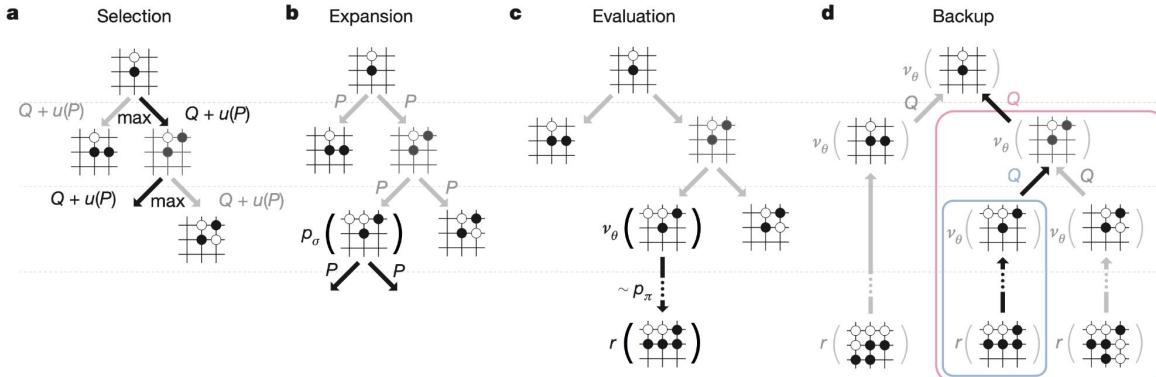
Method	Success Rate (%)		
	Letter	Word	Game
IO	38.7	14	0
CoT	40.6	15.6	1
ToT (ours)	78	60	20
+best state	82.4	67.5	35
-prune	65.4	41.5	5
-backtrack	54.6	20	5

Importance of ...

- Backtrack
- Subtree pruning

ToT: Summary

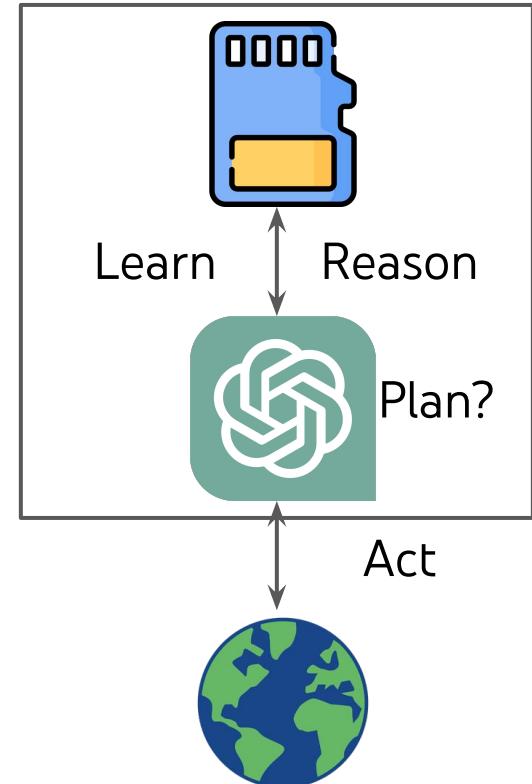
- LM + tree search complement each other
 - LM provides flexible and versatile **heuristics** for search
 - Search augments **control structure** for LLM
- **But how is this related to language agents?**



Language Agent

- We have found some piecewise insightful ideas
 - Language as universal interface across mechanisms

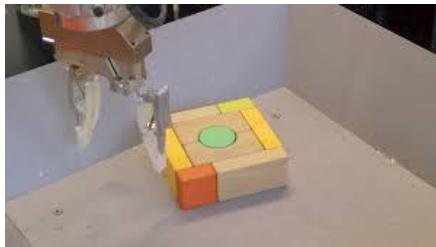
	RL Agent	Language Agent
Action space	Act	Act + Reason
Learning	Parameter update	Language update
Planning	Learned values	LM reasoning for both generation and evaluation



A more systematic framework to synthesize these ideas?

Part II: External Environments

Environment: how to be cheap, fast, yet useful?



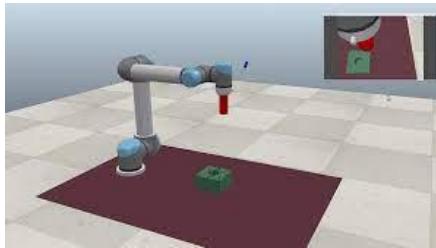
Physical World / Humans



Practical: robots / chatbots



Not Scalable: expensive and slow to collect data



Simulations / Games



Not Practical: sim-to-real is hard



Scalable: free, unlimited interactions



Digital World (Internet, code, software, ...)



Practical: important tasks to automate



Scalable: huge scale, rich complexity, free and fast

Evaluation: how to be cheap, fast, yet high-quality?



Human Evaluation



High-quality (esp. if you're OpenAI): RLHF



Not Scalable (if you're not OpenAI): expensive and slow to collect data

LM Evaluation



Not High-quality (yet): hallucinations



Scalable (somewhat): unlimited, but with costs

$$r = r_{type} \cdot \frac{|U_{att} \cap Y_{att}| +}{|U_{att}|}$$
$$r_{type} \in \{0, 0.1, 0.5, 1\}$$

"Rule-based" Evaluation



High-quality: if we leverage domain priors!



Scalable: unlimited, free, fast

Overview

	1. WebShop 	2. InterCode 	3. Collie 
Task	“Find a t-shirt that...”	“Remove the file that...”	“Write a paragraph that...”
Action	Web actions	Code as action	Text as action
Observation feedback	Webpage	Code execution result	Synthetic constraint satisfaction feedback
Reward feedback	Attribute rule-based	Unit test or file diff	Grammar rule-based

Web Interaction

Table 3
Query
For the last 8 house arrest
In the winter at Rainier Pa

This compar
had a guaran

How can I train the crows in my neighborhood to bring me gifts?

This question does not make sense This question should not be answered

Search results for: how to train crows to bring you gifts Quotes ↗

← how to train crows to bring Find in page ↑ + Add new quote

[How to Make Friends With Crows - PetHelpful](#)
If you did this a few times, your crows would learn your new place, but as I said, I'm not sure if they will follow or visit you there since it's probably not in their territory. The other option is simply to make new crow friends with the crows that live in your new neighborhood.

[Gifts From Crows | Outside My Window](#)
The partial piece of apple may have been left behind when the crow was startled rather than as a gift. If the crows bring bright objects you'll know for sure that it's a gift because it's not something they eat. Brandi Williams says: May 28, 2020 at 7:19 am.

Number of quote tokens left: 463
Number of actions left: 96

Done quoting! Write an answer

(a) Screenshot from the demonstration interface.

♦Question
How can I train the crows in my neighborhood to bring me gifts?

♦Quotes
From Gifts From Crows | Outside My Window (www.birdsoutsidemywindow.org)
Many animals give gifts to members of their own species but crows and other corvids are the only ones known to give gifts to humans.

♦Past actions
Search how to train crows to bring you gifts
Click Gifts From Crows | Outside My Window www.birdsoutsidemywindow.org
Quote
Back

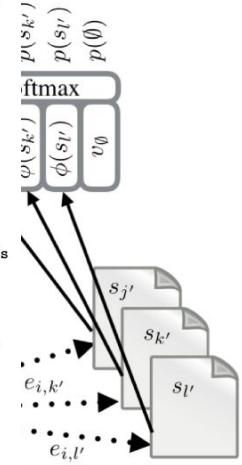
♦Title
Search results for: how to train crows to bring you gifts

♦Scrollbar: 0 - 11
♦Text
[0]How to Make Friends With Crows - PetHelpful|pethelpful.com
If you did this a few times, your crows would learn your new place, but as I said, I'm not sure if they will follow or visit you there since it's probably not in their territory. The other option is simply to make new crow friends with the crows that live in your new neighborhood.

[1]Gifts From Crows | Outside My Window|www.birdsoutsidemywindow.org
The partial piece of apple may have been left behind when the crow was startled rather than as a gift. If the crows bring bright objects you'll know for sure that it's a gift because it's not something they eat. Brandi Williams says: May 28, 2020 at 7:19 am.

♦Actions left: 96
♦Next action

(b) Corresponding text given to the model.



Nogueira et al. End-to-End Goal-Driven Web Navigation
 Nakano et al. WebGPT: Browser-assisted question-answering with human feedback

Mini WWikiNav:WebGPT: RLHF games"mes"

WebShop

A

WebShop search

Instruction:
i'm looking for a small portable folding desk that is already fully assembled; it should have a khaki wood finish, and price lower than 140.00 dollars

portable folding desk khaki wood **1** Search

Back to Search **2** results

Page 1 (Total results: 50) Next >

MENHG Folding Breakfast Tray Table, Efficient Home Laptop Notebook Computer Desk, Portable Writing Study Desk, Sturdy Home Office Table Workstation
\$109.0

KPSP Folding Study Desk Bed Breakfast Serving Tray Table Efficient Home Laptop Notebook Computer Desk Portable Standing Desk for Small Space Bedroom

Description: Product laptop desk. Product design: walnut. Product weight: 4.6 pounds. Material: high quality thick steel pipe, black brushed sheet. Special design: black brushed smooth table top, increase the length and width of the table, it is possible to place the computer and various items. Function: Can be used as computer desk, dining table, bedside table. Product size: 23.6x15.7x11 inches

item-detail

- Large Size
- Wide Application

MENHG Folding Laptop Table Bed Desk **item**

PC Lap Desk with Drawer Book Stand Reading Holder Leg Space Laptop Bed Tray Foldable Lazy Table Breakfast Desk Sofa Small Desk for Small Space

Price: \$100.0 Rating: N/A

Description Overview **Buy Now** **5**

Color: **black** **khaki** **white** **3**

4.1 4.2 **Reward:** 1.0

Why Shopping?

- Scalable **environment**: rich dynamics, scalable items
- Scalable **task**: well-defined problem, possible for automatic reward synthesis
- Interesting **challenges**: multi-modal understanding, decision making, etc.

Construction

Orange: hidden from agents!

- Scrape 1.18M products from amazon.com, text mine attributes

The screenshot shows the Amazon Electronics department page. A black arrow points from the 'Camera & Photo' link in the sidebar to the 'Camera & Photo' section of the main content area. The main content area displays a grid of camera products, with the Kodak PIXPRO Astro Zoom AZ421-BK 16MP Digital Camera being the featured item.

This screenshot shows the product detail page for the Kodak PIXPRO Astro Zoom AZ421-BK 16MP Digital Camera. The page is annotated with numbered callouts:

- KODAK PIXPRO Astro Zoom AZ421-BK 16MP Digital Camera with 42X Optical Zoom and 3" LCD Screen (Black)
- \$174⁹⁵
- Color: Red
- Style: Camera Only
- About this item
 - Powerful 16.1-Megapixel CCD sensor gives you room to enlarge, zoom and crop to content without losing out on life-like clarity or quality.
 - 42x optical zoom and a 24mm wide angle lens Moves you closer to the subject and every frame 1080P
 - Face/Cat/dog detection detects facial features and enhances it for true subject highlight your four-legged Cat and dog friends. Blink/smile detection helps you to capture the moment.
 - Face Beautifier mode and loads of built-in touch-up features allow you to reduce enhance skin tone and the eyes of your portrait subjects instantly.
 - Use the Panorama mode to create a seamless panoramic picture to 180-degree eas the camera across your subject. Continuous shooting
 - The az421 takes a standard SD/SDHC memory card that is at least Class 4 and no less than 32GB
- Product Description

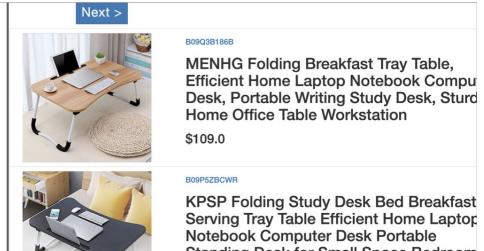
An orange arrow points from the 'About this item' section to the right, pointing towards the text 'Optical zoom', 'LCD screen', 'Memory card', and 'Digital camera'.

7
Optical zoom
LCD screen
Memory card
Digital camera

Construction

Orange: hidden from agents!

- Scrape 1.18M products from amazon.com, text mine attributes
- Build synthetic website with aligned text interface



The screenshot shows a product listing for a 'MENHG Folding Breakfast Tray Table'. It includes a main image of the desk and a smaller image below it. The product title is 'MENHG Folding Breakfast Tray Table, Efficient Home Laptop Notebook Computer Desk, Portable Writing Study Desk, Sturdy Home Office Table Workstation'. The price is \$109.0. Below the main image is another product listing for a 'KPSP Folding Study Desk Bed Breakfast Serving Tray Table Efficient Home Laptop Notebook Computer Desk Portable Standing Desk for Small Space Bedroom'.

HTML mode

Simple mode

B Instruction:
I'm looking for a small portable folding desk
that is already fully assembled [...]
[btn] Back to Search [/btn]
Page 1 (Total results: 50) [btn] Next [/btn]
[btn] MENHG Folding Breakfast Tray [...] [/btn]
\$109.0
[btn] KPSP Folding Study Desk Bed [...] [/btn]

Construction

Orange: hidden from agents!

- Scrape 1.18M products from amazon.com, text mine attributes
- Build synthetic website with aligned text interface
- **Human** instructions & **Automatic** reward via product attributes/options/prices/types

Goal product (U)



Fujifilm X-T1 16 MP Mirrorless Digital Camera with 3.0-Inch LCD (Body Only) (Graphite Silver & Weather Resistant) (Renewed)
Price: \$904.95

Attributes

certified refurbished
water resistant
high performance

Human →

Goal Instruction

"i'd like to find a digital camera that's water resistant. the color needs to be graphite silver and i want the configuration to be the international version, and price under 1000 dollars"

- Goal Attributes: [water resistant]
- Goal Options: [graphite silver, international version]

Construction

Orange: hidden from agents!

- Scrape 1.18M products from amazon.com, text mine attributes
- Build synthetic website with aligned text interface
- **Human** instructions & **Automatic** reward via product attributes/options/prices/types

Purchased product (Y)



Olympus E-PL5 16MP Mirrorless Digital Camera with 3-Inch LCD, Body Only (White) (Old Model)
Price: \$139.79

color	black	silver	white
style	body only	w 14-42mm	double zoom kit
configuration	base	international version	

Attributes
touch screen
digital camera

Goal Instruction

"i'd like to find a digital camera that's water resistant. the color needs to be graphite silver and i want the configuration to be the international version, and price under 1000 dollars"

- Goal Attributes: [water resistant]
- Goal Options: [graphite silver, international version]



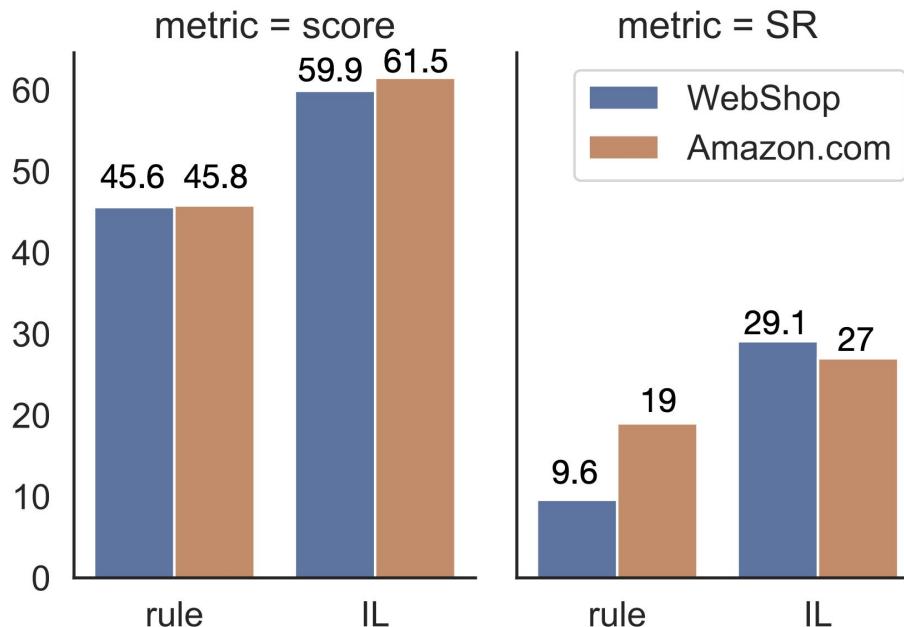
Automatic Reward Calculation

$$r = r_{type} \cdot \frac{|U_{att} \cap Y_{att}| + |U_{opt} \cap Y_{opt}| + \mathbf{1}[y_{price} \leq u_{price}]}{|U_{att}| + |U_{opt}| + 1}$$
$$r_{type} \in \{0, 0.1, 0.5, 1\}$$

- 1) attribute matching = 0/1
- 2) option matching = 2/2
- 3) price matching = 1/1
- 4) Type matching = 1.0

$$r = 1.0 \cdot \frac{0 + 2 + 1}{1 + 2 + 1} = 0.75$$

Sim-to-real transfer (Amazon/eBay)

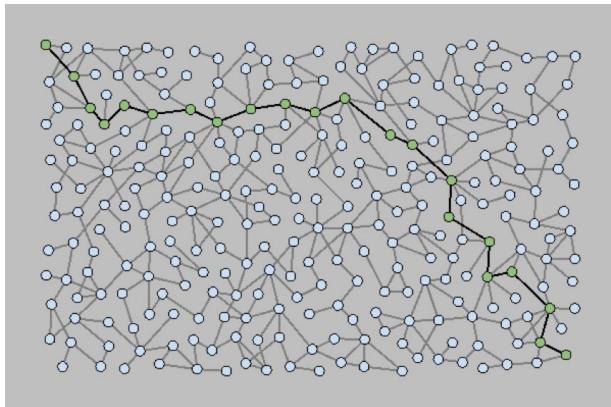


Learned policy generalizes to different search engine and products

WebShop: Summary

- First large-scale, realistic web interaction benchmark
 - Trending in 2023: WebArena, Mind2Web, ...
- **Synthetic website** provides controlled development, transfers to real websites
- **Task priors** (self/model-supervision) provides scalable reward

(Individual) Human reward finetuning on top of synthetic reward pre-training?

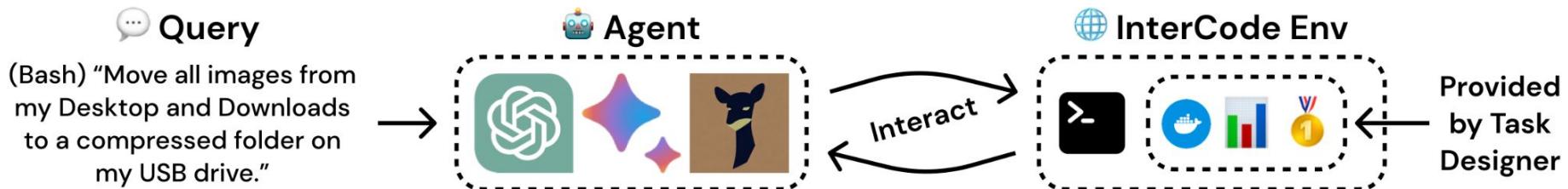


Text mining
touch screen
digital camera

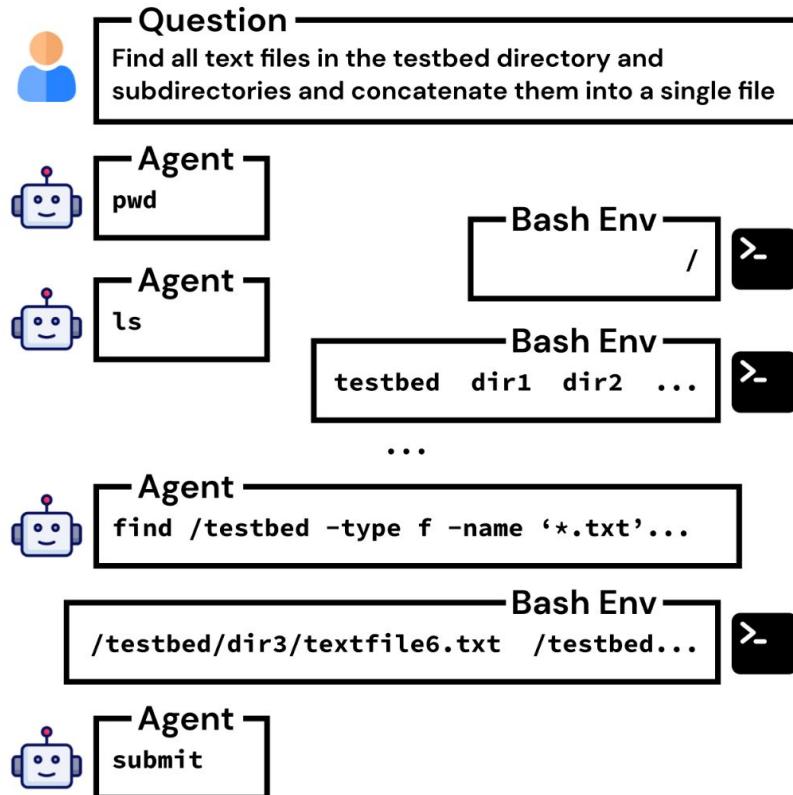
→
Sentiment analysis
Users find it good..
Summarization
This is a digital...
Image Detection
Optical zoom...

Code Interaction

- Static NL2Code benchmarks: HumanEval, Spider, NL2Bash, ...
- But humans code in a fundamentally interactive manner!
- Some interactive/execution-based methods, but no standard benchmark

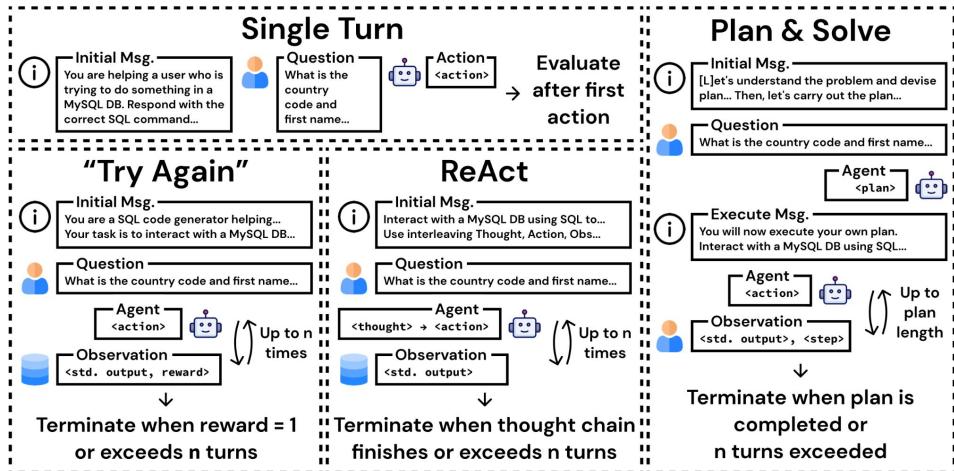


InterCode Setup



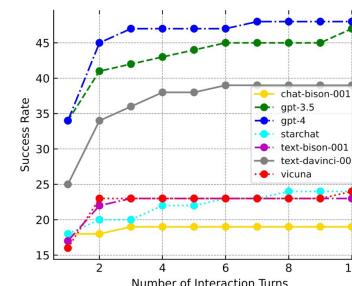
- Standard RL env:
 - **Environment:** Docker-based Python/SQL/bash terminals
 - **Action:** code command
 - **Observation:** execution result
- Benefits
 - Safe and reproducible
 - Unlock new tasks (e.g. CTF)
 - Unlock new evaluations (e.g. Bash)
 - Unlock new methods (e.g. Plan-and-solve)

InterCode: new methods

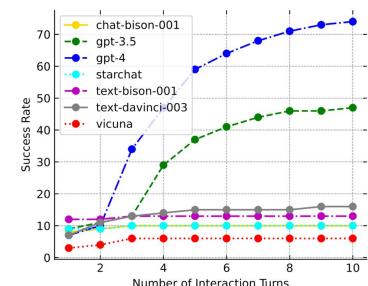


Try Again ($n = 10$)			ReAct ($n = 10$)			Plan & Solve			
SR	Turns	Error %	SR	Turns	Error %	SR	Turns	Error %	
SQL	47.3	7.25	46.4	58.7	5.30	6.94	49.1	4.29	16.2
Bash	46.5	6.15	24.9	20.5	4.40	20.4	28.0	6.65	53.3

- Interactive >> seq2seq
- Different interactive methods have different tradeoffs
- Large room for improvement

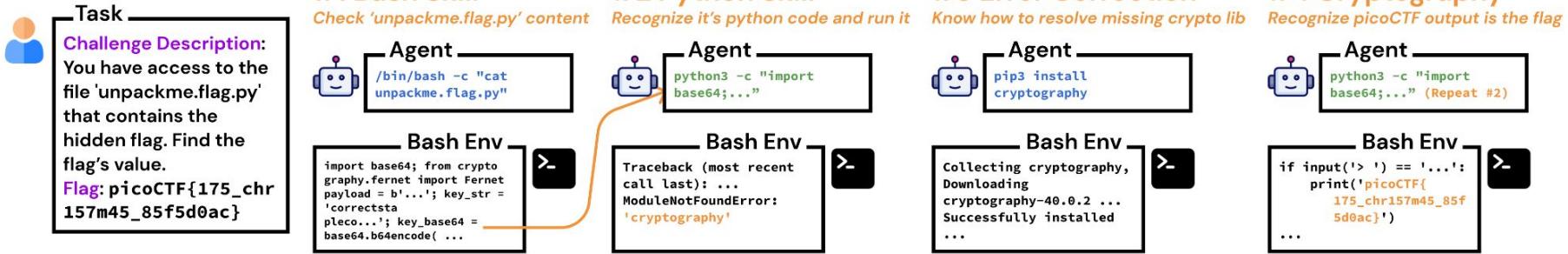


(a) Success rate vs. turns for InterCode-Bash



(b) Success rate vs. turns for InterCode-SQL

Future: Coding -> Software engineering?



Constrained Text Generation

- A traditional and important NLP (seq2seq) task: constraints -> text
- **Prior benchmarks:** fixed constraint type and too simple for LLMs
 - “Generate a sentence with dog, catch, happy.”
- **Collie’s goal:**
 - Diverse and arbitrarily hard constraints for LLMs, yet guaranteed to be solvable
 - Automatic task construction + evaluation, without human efforts!
 - Challenge language understanding/generation, semantic planning, logical/arithmetic reasoning, ...
- **Core idea:** leverage the infinite expressivity of **grammar**

Collie: Grammar

$$S \rightarrow (\text{level}(\xi) = \ell) \wedge M \quad (\text{constraint specification}) \quad (1)$$

$$M \rightarrow C \mid C \wedge M \mid C \vee M \quad (\text{multi-constraint}) \quad (2)$$

$$C \rightarrow \text{count}(T, \ell, v_{\text{str}} \mid \ell') \oplus v_{\text{num}} \mid \text{pos}(T, \ell, v_{\text{num}}) \circ v_{\text{str}} \quad (\text{base-constraint}) \quad (3)$$

$$T \rightarrow \xi \mid \text{pos}(T, \ell, v_{\text{num}}) \quad (\text{text}) \quad (4)$$

$$\ell \rightarrow \text{char} \mid \text{word} \mid \text{sentence} \mid \text{paragraph} \mid \text{passage} \quad (\text{level}) \quad (5)$$

$$\circ \rightarrow = | \neq \quad \oplus \rightarrow = | \neq | > | < | \leq | \geq \quad (\text{relation}) \quad (6)$$

$$v_{\text{str}} \in \Sigma^* \quad v_{\text{num}} \in \mathbb{Z} \quad (\text{value}) \quad (7)$$

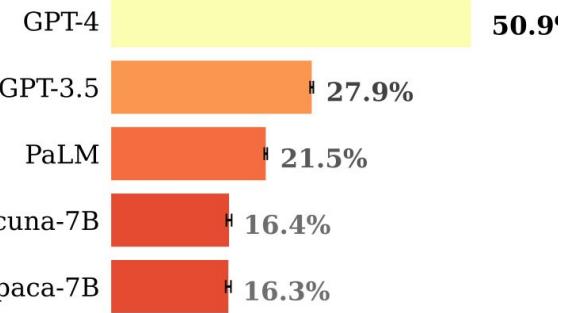
- **Few core concepts:** count, pos, level
- But compositionality yields power
- Easily extensible (e.g. POS, sentiment, topic, ...)

Collie-v1: 2,080 Constraints across 13 Types

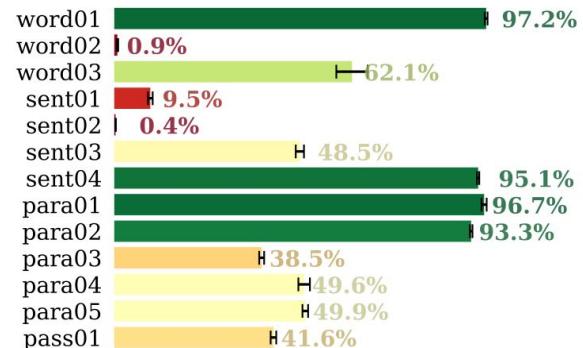
ID	Example instruction	Multi-constraint M
word01	Generate a word with at least 15 letters.	$\text{count}(\xi, \text{char}, \text{word}) \geq 15$
word02	Generate a word with 10 letters, where letter 1 is ‘s’, letter 3 is ‘r’, letter 9 is ‘e’.	$\text{count}(\xi, \text{char}, \text{word}) = 10 \wedge \text{pos}(\xi, \text{char}, 1) = \text{'s}' \wedge \text{pos}(\xi, \text{char}, 3) = \text{'r'} \wedge \text{pos}(\xi, \text{char}, 9) = \text{'e'}$
word03	Generate a word with at most 10 letters and ends with “r”.	$\text{count}(\xi, \text{char}, \text{word}) \leq 10 \wedge \text{pos}(\xi, \text{char}, -1) = \text{'r'}$
sent01	Please generate a sentence with exactly 82 characters. Include whitespace into your character count.	$\text{count}(\xi, \text{char}, \text{sentence}) = 82$
sent02	Generate a sentence with 10 words, where word 3 is “soft” and word 7 is “beach” and word 10 is “math”.	$\text{count}(\xi, \text{word}, \text{sentence}) = 10 \wedge \text{pos}(\xi, \text{word}, 3) = \text{"soft"} \wedge \text{pos}(\xi, \text{word}, 7) = \text{"beach"} \wedge \text{pos}(\xi, \text{word}, 10) = \text{"math"}$
sent03	Generate a sentence with at least 20 words, and each word less than six characters.	$\text{count}(\xi, \text{word}, \text{sentence}) \geq 20 \wedge \text{count}(\xi, \text{char}, \text{word}) \leq 6$
sent04	Generate a sentence but be sure to include the words “soft”, “beach” and “math”.	$\text{count}(\xi, \text{word}, \text{'soft'}) > 0 \wedge \text{count}(\xi, \text{word}, \text{'beach'}) > 0 \wedge \text{count}(\xi, \text{word}, \text{'math'}) > 0$
para01	Generate a paragraph where each sentence begins with the word “soft”.	$\text{pos}(\text{pos}(\xi, \text{sentence}, 1), \text{word}, 1) = \text{'soft'} \wedge \text{pos}(\text{pos}(\xi, \text{sentence}, 2), \text{word}, 1) = \text{'soft'} \wedge \dots$
para02	Generate a paragraph with at least 4 sentences, but do not use the words “the”, “and” or “of”.	$\text{count}(\xi, \text{sentence}, \text{paragraph}) \geq 4 \wedge \text{count}(\xi, \text{word}, \text{'the'}) = 0 \wedge \text{count}(\xi, \text{word}, \text{'and'}) = 0 \wedge \text{count}(\xi, \text{word}, \text{'of'}) = 0$
para03	Generate a paragraph with exactly 4 sentences, each with between 10 and 15 words.	$\text{count}(\xi, \text{sentence}, \text{paragraph}) = 4 \wedge \text{count}(\xi, \text{word}, \text{sentence}) \geq 10 \wedge \text{count}(\xi, \text{word}, \text{sentence}) \leq 15$
para04	Generate a paragraph with at least 3 sentences, each with at least 15 words.	$\text{count}(\xi, \text{sentence}, \text{paragraph}) \geq 3 \wedge \text{count}(\xi, \text{word}, \text{sentence}) \geq 15$
para05	Generate a paragraph with 2 sentences that end in “math” and “rock” respectively.	$\text{count}(\xi, \text{sentence}, \text{paragraph}) = 2 \wedge \text{pos}(\text{pos}(\xi, \text{sentence}, 1), \text{word}, -1) = \text{"math"} \wedge \text{pos}(\text{pos}(\xi, \text{sentence}, 2), \text{word}, -1) = \text{"rock"}$
pass01	Generate a passage with 2 paragraphs, each ending in “I sit.” and “I cry.” respectively.	$\text{count}(\xi, \text{paragraph}, \text{passage}) = 2 \wedge \text{pos}(\text{pos}(\xi, \text{paragraph}, 1), \text{sentence}, -1) = \text{"I sit."} \wedge \text{pos}(\text{pos}(\xi, \text{paragraph}, 2), \text{sentence}, -1) = \text{"I cry."}$

(a)

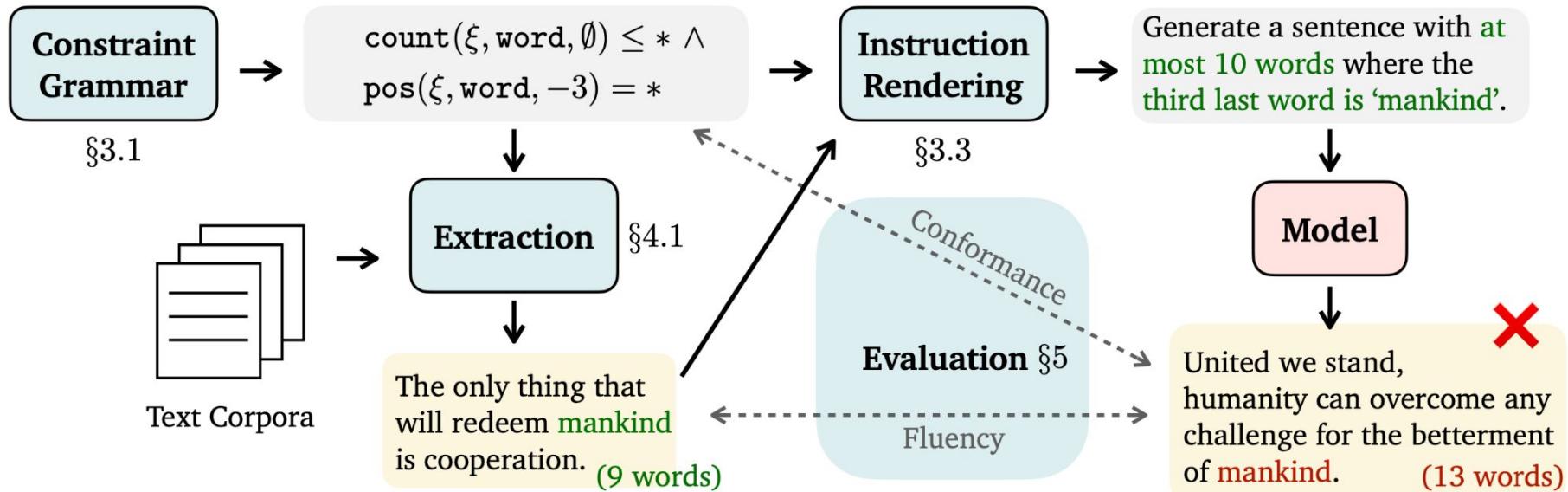
Zero-shot performance



Constraint satisfaction rate - GPT-4



Task Construction is Fully Automatic



- Human just specify constraint types
- Collie automatically extracts constraint “values” from corpora
- Rule-based instruction rendering and text evaluation (extensible)

Feedback helps!

SH

generate a word with 10 letters, where letter 1 is "s", letter 3 is "r", letter 9 is "e"



Sure, here's a word that meets your criteria: "strenuous".

SH

your word only has 9 letters, and letter 9 is "s", not "e".

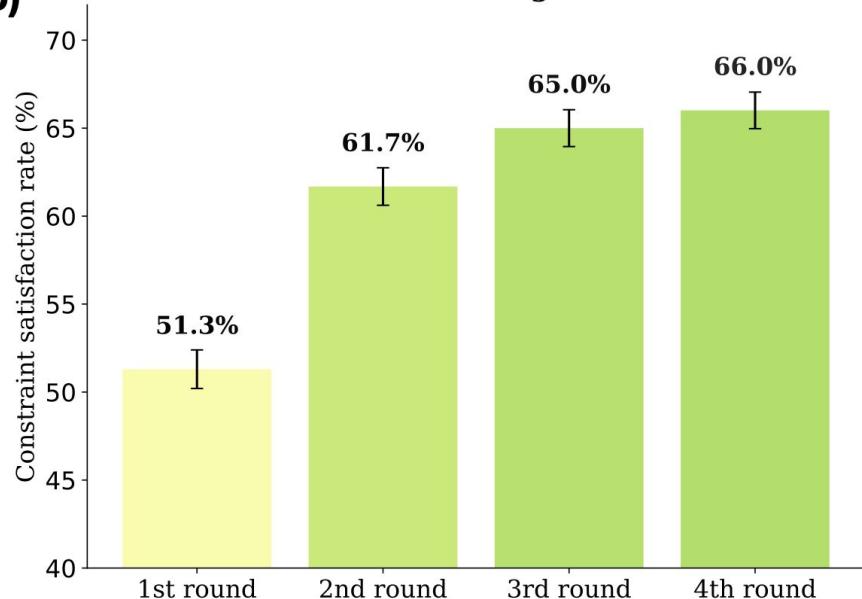


I apologize for the mistake. Here's a correct word that meets your criteria: "strategies".

Many other cool findings in the paper!

(b)

GPT-4 interactive generation

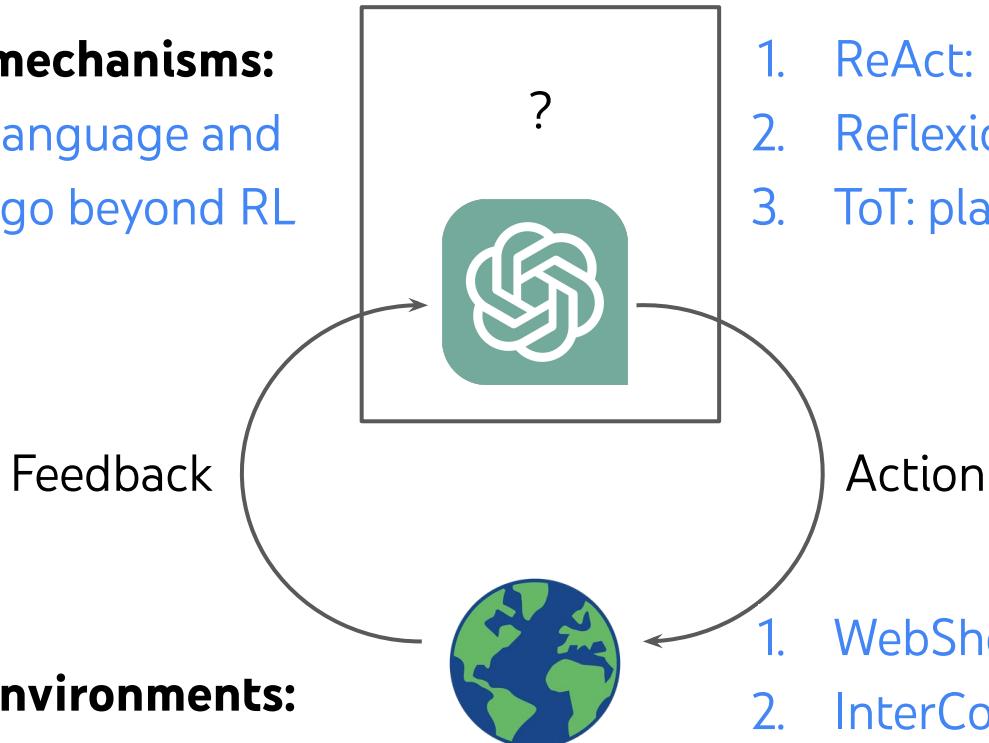


Thanks!

Internal mechanisms:

Leverage language and reasoning, go beyond RL

1. ReAct: reasoning
2. Reflexion: learning
3. ToT: planning



External environments:

Digital domains, scalable and faithful evaluation

1. WebShop: web
2. InterCode: code
3. Collie: logic