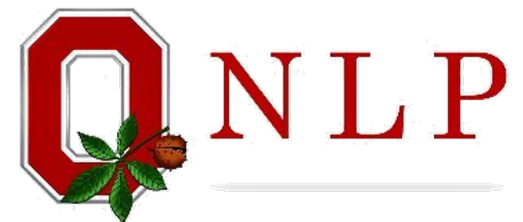


# Language agents: a critical evolutionary step of artificial intelligence

Yu Su

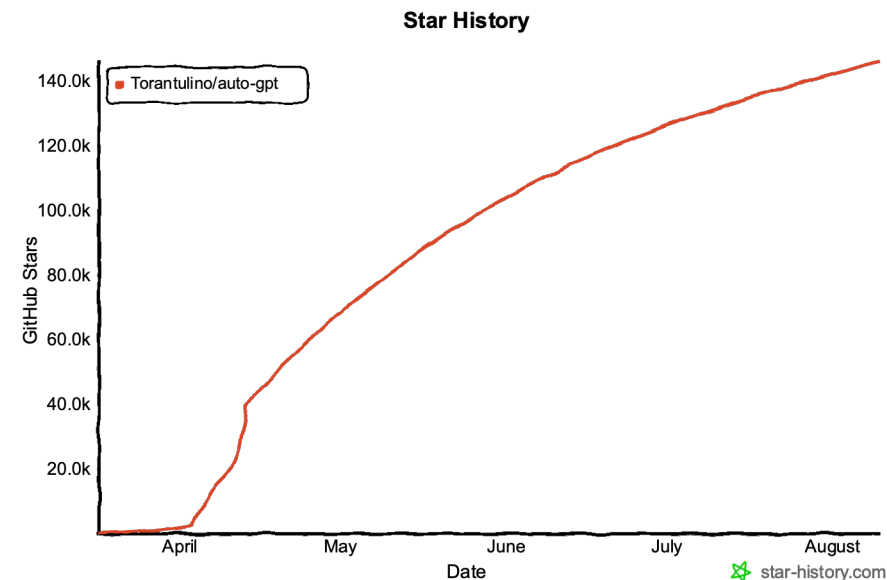
The Ohio State University



# Language agents

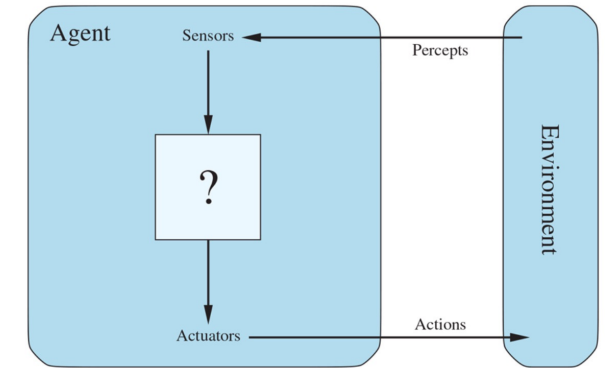
- **Definition:** autonomous agents, usually powered by large language models, that can follow language instructions to carry out diverse and complex tasks in real-world or simulated environments.
- **Examples:** [Auto-GPT](#), [GPT-Engineer](#), [Voyager](#), [RT-2](#), and many others

Probably the most heated thread in AI right now with massive public interests, e.g., [Auto-GPT](#) has received over 145k stars on GitHub within merely 4 months, making it the [fastest growing repository](#)



# But why?

- The concept of agent has been introduced in AI since its dawn. It's the first concept we teach in AI 101. What's different this time around?
- I argue that the most fundamental change is the *capability of using language*.
- Contemporary AI agents use language as a vehicle for both thought and communication, a trait that was unique to humans.



Russel & Norvig, 2020

## Chain-of-Thought Prompting

Model Input

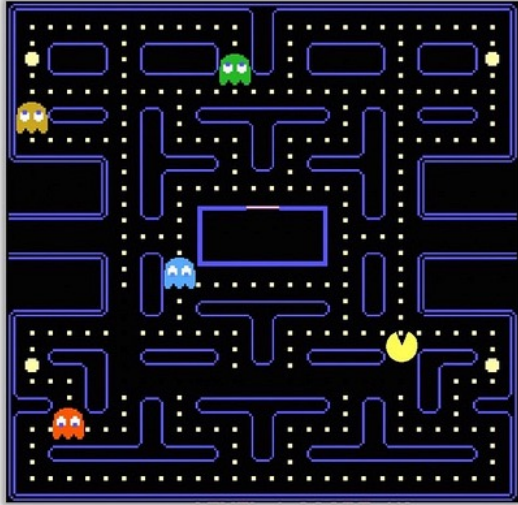
Q: Roger has 5 tennis balls. He buys 2 more cans of tennis balls. Each can has 3 tennis balls. How many tennis balls does he have now?

Therefore, these contemporary AI agents capable of using language for thought and communication should be called “**language agents**,” for language being their most salient trait.

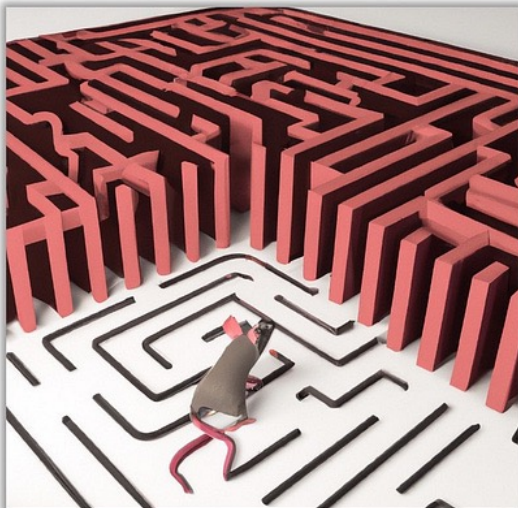
0 months 50 months 100 months 150 months 200 months 250 months

bought 6 more apples, so they have  $3 + 6 = 9$ . The answer is 9. ✓

# Evolution of biological intelligence: an analogy



Artificial Intelligence



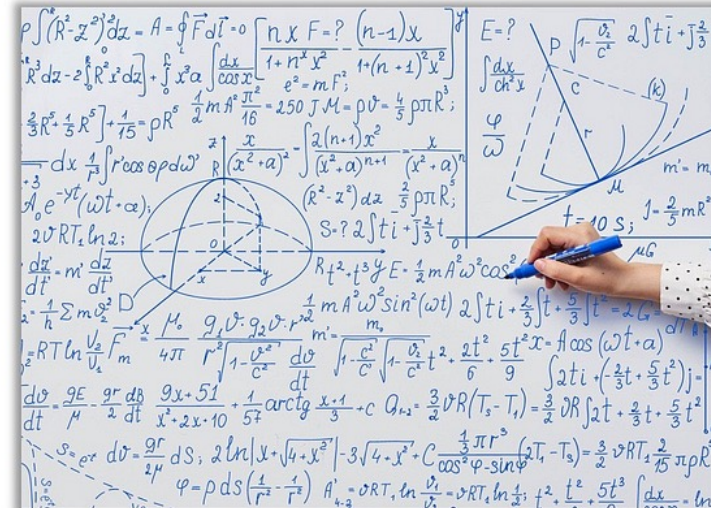
Biological Intelligence



OS If we have room-temperature superconductor, what would it mean for artificial intelligence?

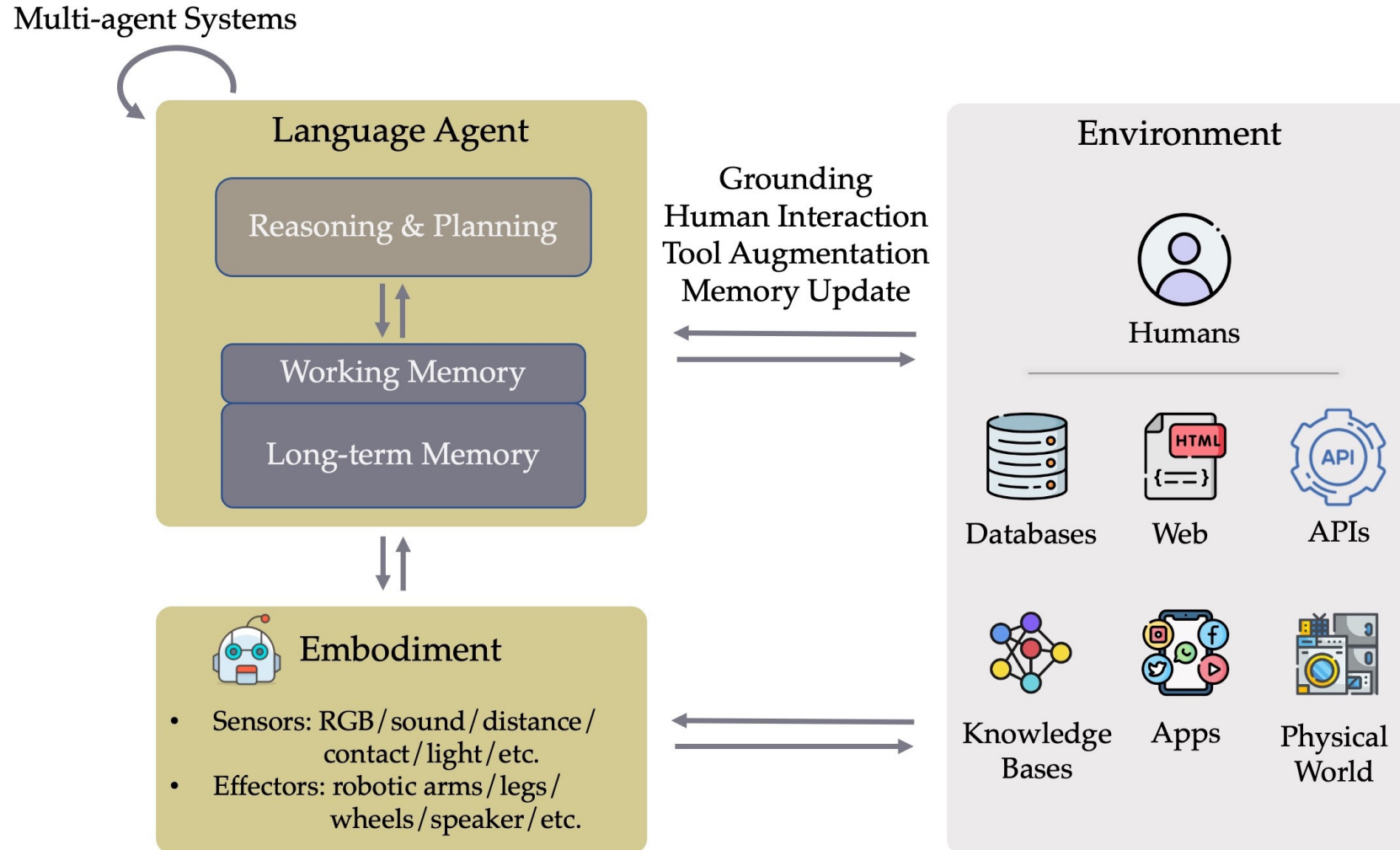
The discovery of a room-temperature superconductor would have profound implications for various fields, including artificial intelligence (AI). Here's how it could impact AI and related technologies:

- Energy Efficiency:** Superconductors have zero electrical resistance, which means electrical currents can flow without any energy loss. This could lead to the development of highly energy-efficient hardware for AI computations. As AI models and algorithms become more complex, they require more computational power, which in turn requires more energy. Room-temperature superconductors could significantly reduce the energy consumption of AI data centers.
- Faster Computation:** Reduced resistance means faster electrical signal transmission. This could lead to faster computation speeds, allowing AI models to process data and make decisions more rapidly.
- Quantum Computing:** Superconductors are essential components of many quantum computers, which use quantum bits (qubits) instead of the binary bits used in classical





# Language agents: a conceptual framework



# Memory

- Animals store memory through *synaptic plasticity*; artificial neural networks also store information in their weights ( $\approx$  synaptic strength)
- In-context learning  $\approx$  working memory
- Long-term memory: LLMs' *parametric memory* or *vector database*?
- Why LLMs can understand and generate language? LLMs may have internalized a compact “linguistic representation” by compressing the pre-training corpus, similar to how humans learn and memorize
- But can an LLM truly “understand” a concept (e.g., “apple”) without actually “seeing” the physical object?

# Textual twin world theory

OS

what is fruit



A fruit is the mature ovary of a flowering plant, usually containing seeds. In botanical terms, a fruit develops from the fertilized ovary of a flower after pollination and is often involved in the process of seed dispersal, allowing plants to spread their seeds and reproduce.

OS

what is an apple



An apple is a type of fruit that is commonly grown on apple trees (*Malus domestica*). It is one of the most widely cultivated and consumed fruits in the world. Apples come in a variety of colors, including red, green, and yellow, and they can range in flavor from sweet to tart. The fruit has a round shape with a thin, edible skin and a crisp, juicy flesh inside.

OS

what is the color red



Red is a color in the visible light spectrum, and it is often described as one of the primary colors along with blue and green. It is a color that is associated with a range of meanings and emotions, including passion, love, energy, warmth, and intensity. In terms of wavelengths, red light has longer wavelengths than blue or green light.

OS

what is round shape



A round shape is a two-dimensional geometric shape that is characterized by having all points on its boundary equidistant from its center. It is symmetrical and lacks corners or edges. A circle is a classic example of a round shape in two dimensions.

# Tool augmentation

- Solely relying on an LLM's static parameters can only do limited things. Tool augmentation brings unlimited possibilities.
- Three main purposes of tool augmentation
  - Complement language agents with **specialized capabilities** they may not have, e.g., high-precision calculation and routing on a map
  - Provide **up-to-date information**, e.g., retriever and search engine
  - Enable language agents to **take actions** in real-world environments
- Two types of tools: read-only vs. state-changing (i.e., w/ side effects)
- **Robustness** and **flexibility** in using tools is key. None of the existing systems, including ChatGPT Plugins, is sufficiently robust.
  - Tools w/ side effects are riskier and require higher robustness.

# Reasoning and planning

- Reasoning is a continuum, not black and white. There has been too much evidence for LLMs' various reasoning capabilities. Simply arguing that LLMs cannot reason at all may not be constructive discussion.
- Traditionally, reasoning and planning are relatively separate topics in AI.
  - Reasoning: deductive (e.g., logic-based), inductive (e.g., ML)
  - Planning: Given start and goal states, an action space, find (the optimal) solutions
- But now they are getting blended in language agents, consider, e.g., an embodied language agent acting in a partially-observable environment.
- New reasoning algorithms (e.g., chain/tree-of-thought) have emerged to unleash LLMs' capability of using language for thought, but the truly transformative reasoning algorithms for language agents are probably yet to come



# Grounding

- Each environment is a unique context that provides possibly different interpretations of natural language, which brings the challenge of *grounding*, i.e., linking of (natural language) concepts to contexts (Chandu et al., 2021)
- Two types of grounding are central to language agents
  - **#1: Ground natural language to an environment**, e.g., mapping an utterance to the right API call, SQL query, or robot plan
  - **#2: Ground an agent's generation and decisions to its own context**, including external information generated by tools
- Both remain challenging for current language agents
  - Some good news for #2? [Our recent work](#) finds that LLMs are highly receptive to external evidence, if we present it in a coherent and convincing way

Xie et al., 2023. [Adaptive Chameleon or Stubborn Sloth: Unraveling the Behavior of Large Language Models in Knowledge Clashes](#). arxiv preprint.

# For the rest of the talk

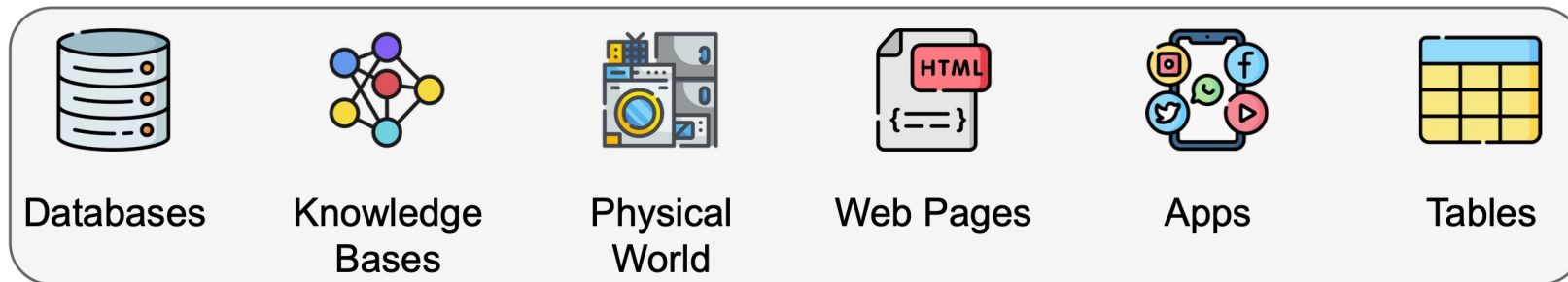
- **Don't Generate, Discriminate: A Proposal for Grounding Language Models to Real-World Environments**  
Yu Gu, Xiang Deng, Yu Su  
*ACL 2023 (Outstanding Paper Award)*
- **Mind2Web: Towards a Generalist Agent for the Web**  
Xiang Deng, Yu Gu, Boyuan Zheng, Shijie Chen, Samuel Stevens, Boshi Wang, Huan Sun, Yu Su  
*Under Review*
- **LLM-Planner: Few-Shot Grounded Planning for Embodied Agents with Large Language Models**  
Chan Hee Song, Jiaman Wu, Clayton Washington, Brian M. Sadler, Wei-Lun Chao, Yu Su  
*ICCV 2023*

# Grounded language understanding

Given a natural language utterance  $u$  and a target environment  $E$

$$\pi: (u, E) \rightarrow p, \text{ s.t. } \llbracket u \rrbracket_E = \llbracket p \rrbracket_E$$

Where  $p$  is a plan/program in a formal language, and  $\llbracket \cdot \rrbracket_E$  is the denotation



# Grounded language understanding

Given a natural language utterance  $u$  and a target environment  $E$

$$\pi: (u, E) \rightarrow p, \text{ s.t. } \llbracket u \rrbracket_E = \llbracket p \rrbracket_E$$

Where  $p$  is a plan/program in a formal language, and  $\llbracket \cdot \rrbracket_E$  is the denotation

$u$ : *What is the latest released computer emulator developed in Java?*

$p$ : (ARGMAX (AND ComputerEmulator  
(JOIN LanguagesUsed Java))  
LatestReleaseDate)



Knowledge  
Bases

# Grounded language understanding

Given a natural language utterance  $u$  and a target environment  $E$

$$\pi: (u, E) \rightarrow p, \text{ s.t. } \llbracket u \rrbracket_E = \llbracket p \rrbracket_E$$

Where  $p$  is a plan/program in a formal language, and  $\llbracket \cdot \rrbracket_E$  is the denotation

$u$ : *Bring me a cup of coffee*

$p$ : [turn left, move forward, pick up cup, turn around, move forward, ..., put cup in coffee maker, toggle coffee maker, ...]



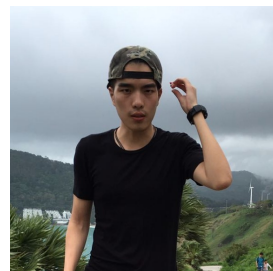
Physical  
World



# Pangu: A Unified Framework for Grounded Language Understanding

Yu Gu, Xiang Deng, Yu Su

The Ohio State University



QUIZ  
TIME!

# Q1: Find the right program over a KB

**Question:** Who has ever coached an ice hockey team in Canada?

## **Program:**

- A. (AND cricket.cricket\_coach (JOIN cricket.cricket\_team.coach\_inv (JOIN sports.sports\_team.location Canada)))
- B. (AND ice\_hockey.hockey\_coach (JOIN ice\_hockey.hockey\_team.coach\_inv (JOIN sports.sports\_team.location Canada)))
- C. (AND ice\_hockey.hockey\_team (JOIN sports.sports\_team.location Canada))



# Q2: Write the corresponding KB program

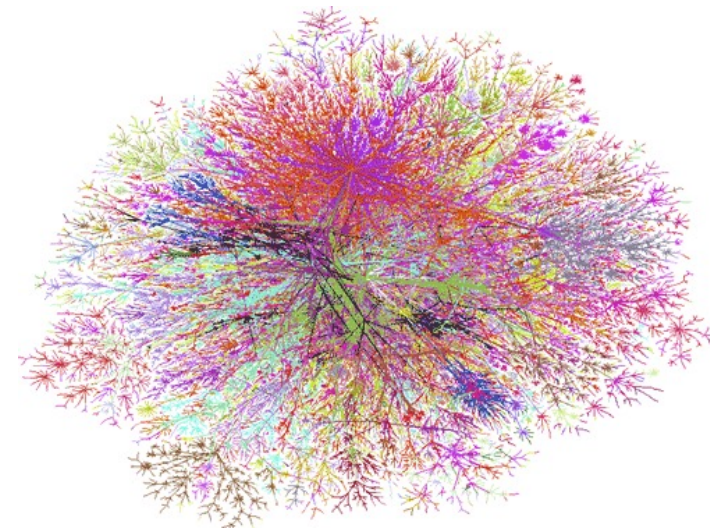
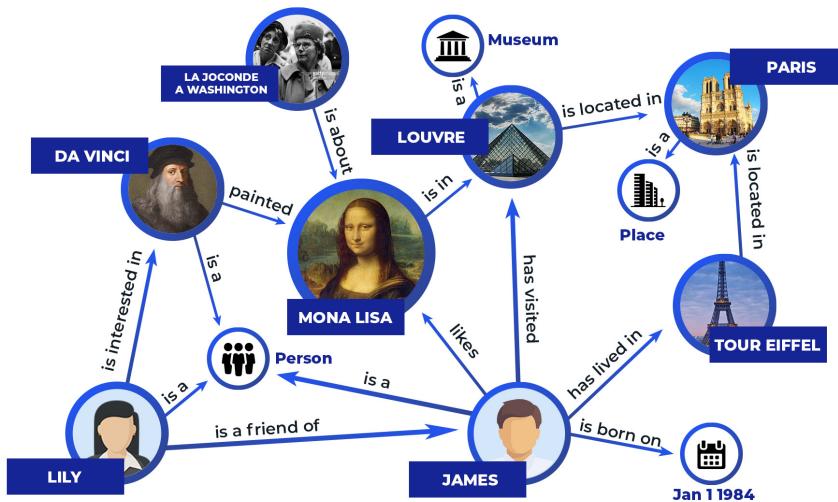
**Question:** What's the classification of the M10 engine?

**Program:**

```
(AND automotive.engine_type (JOIN automotive.engine_type.used_in M10))
```

# Why is Q2 harder?

- 1 You need to learn the grammar
- 2 You need to know the environment specifics





# Key message

**Directly generating plans (programs)  
may not be the optimal way of using  
LMs for grounded language  
understanding**

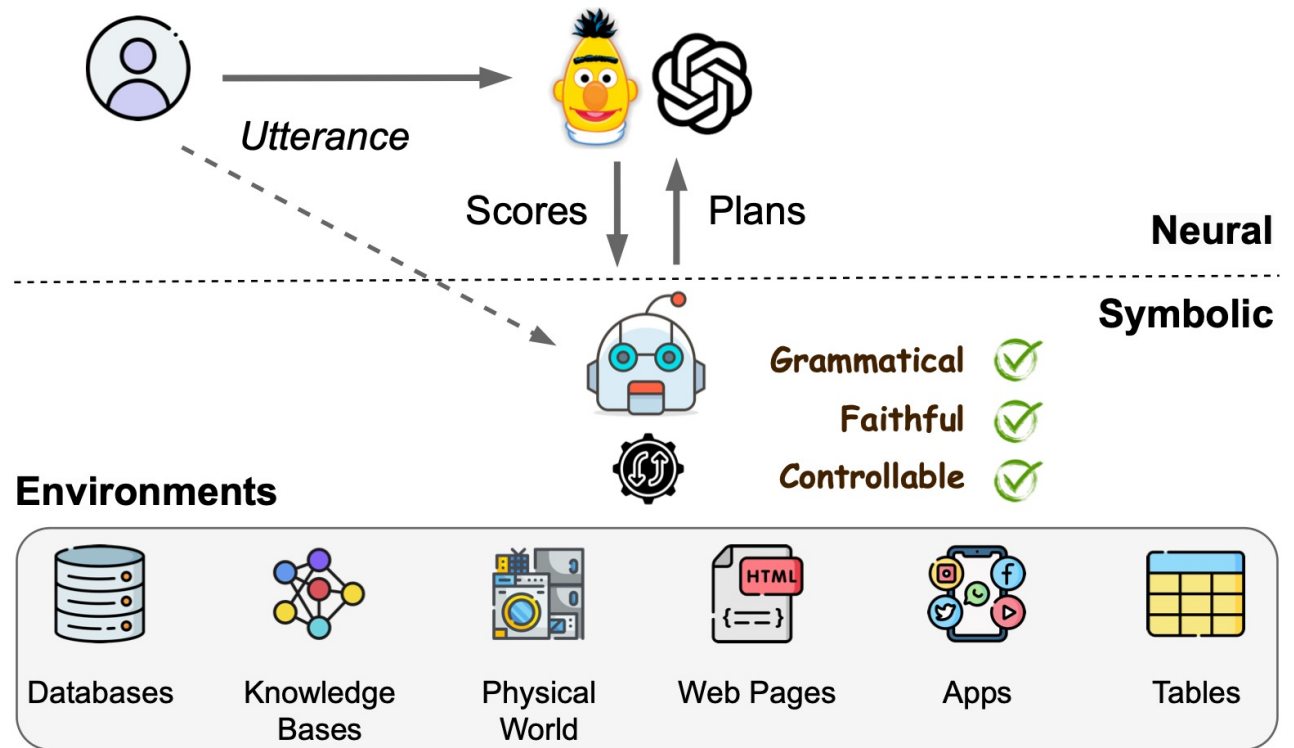
# Pangu:

**A unified framework that models grounded language understanding as a discrimination task**

# Our proposal: Pangu

## Goals:

- Allow LMs to focus on discrimination
- Generic for different tasks



A symbolic agent searches the environment to propose valid candidate plans, while a neural LM scores the plans to guide the search process

# Algorithmic definition

---

## Algorithm 1: PANGU

---

```
1 Input: utterance  $q$ , initial plans  $P_0$ , environment  $E$ 
2  $t \leftarrow 1$ ;
3 while True do
4   /* AGENT PROPOSES PLANS */
5    $C_t \leftarrow \mathbf{Candidate-Plans}(P_{t-1}, E)$ 
6   /* LM SCORES AND PRUNES PLANS */
7    $P_t \leftarrow \mathbf{Top-K}(q, C_t)$ 
8   if Check-Termination() = True then
9     return top-scored plan
10   $t \leftarrow t + 1$ 
```

Initialization of search

Propose candidate plans from the environment

Rank candidate plans using a language model

Repeat until the termination condition is met

# Instantiation for KBQA

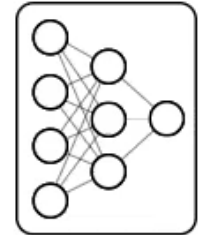


## Testbed:

- KBQA
  - 45M entities
  - 3B facts

## LMs:

- BERT
- T5
- Codex





# New SoTA for KBQA

Prior Art	78.7
Pangu w/ BERT-base	79.9
Pangu w/ T5-base	79.9
Pangu w/ T5-3B	<b>81.7</b>

Prior Art	34.3
Pangu w/ BERT-base	52.0
Pangu w/ T5-base	53.3
Pangu w/ T5-3B	<b>62.2</b>

Prior Art	78.8
Pangu w/ BERT-base	77.9
Pangu w/ T5-base	77.3
Pangu w/ T5-3B	<b>79.6</b>

F1 on GrailQA  
(i.i.d. + non-i.i.d., ~45K  
training examples)

F1 on GraphQuestions  
(non-i.i.d., ~2K training  
examples)

F1 on WebQSP  
(i.i.d., ~3K training  
examples)

## Findings:

- 1 Particularly strong performance for non-i.i.d. generalization
- 2 Stable gain from increased model size

# In-context learning with LLMs

Prior Art	<b>78.7</b>
Codex 10-shot	48.9
Codex 100-shot	53.3
Codex 1000-shot	56.4

Prior Art	34.3
Codex 10-shot	42.8
Codex 100-shot	43.3
Codex 1000-shot	<b>44.3</b>

Prior Art	<b>78.8</b>
Codex 10-shot	45.9
Codex 100-shot	54.5
Codex 1000-shot	68.3

F1 on GrailQA  
(i.i.d. + non-i.i.d., ~45K  
training examples)

F1 on GraphQuestions  
(non-i.i.d., ~2K training  
examples)

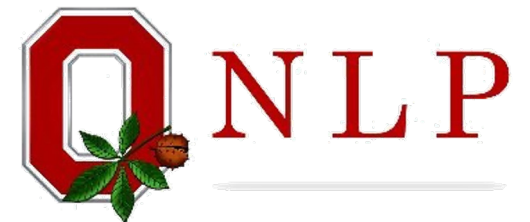
F1 on WebQSP  
(i.i.d., ~3K training  
examples)

## Findings:

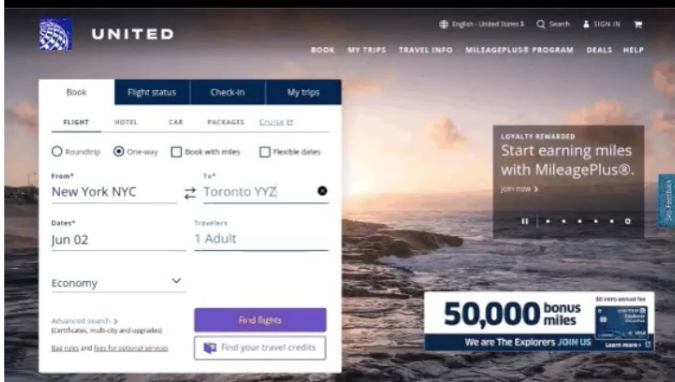
- 1 SoTA performance on GraphQ with only 10 training examples
- 2 Marginal gain from more training data for non-i.i.d.

# Mind2Web: Towards a Generalist Agent for the Web

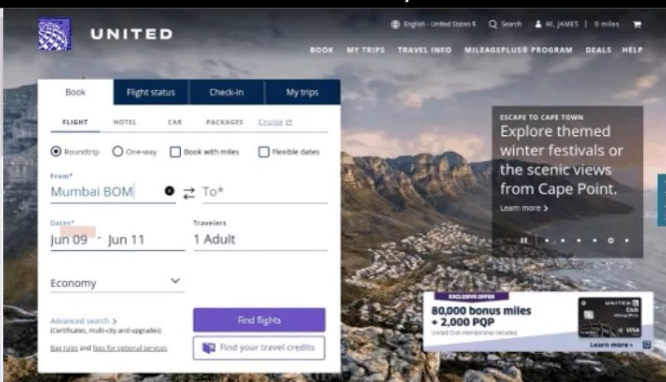
Xiang Deng, Yu Gu, Boyuan Zheng,  
Shijie Chen, Samuel Stevens, Boshi Wang,  
Huan Sun, Yu Su



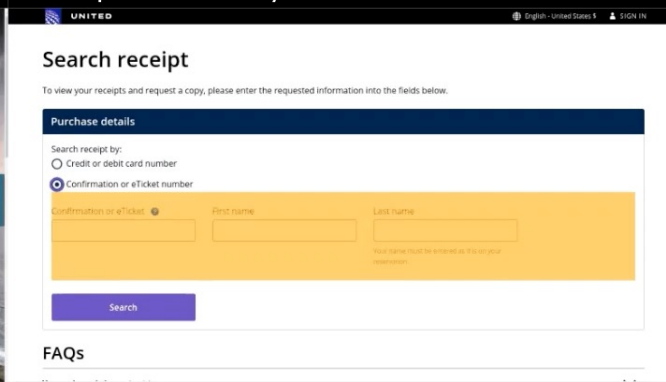
(a) Find one-way flights from New York to Toronto.



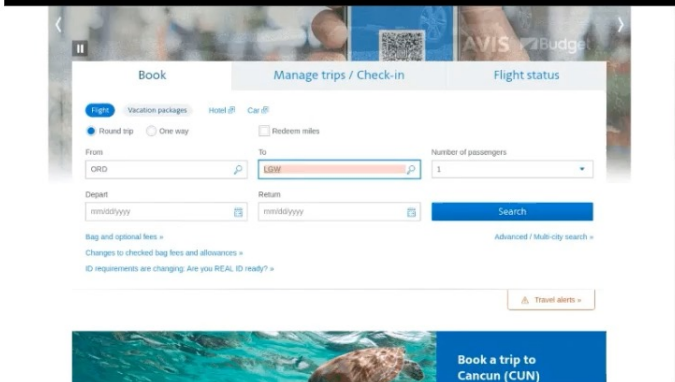
(b) Book a roundtrip on July 1 from Mumbai to London and vice versa on July 5 for two adults...



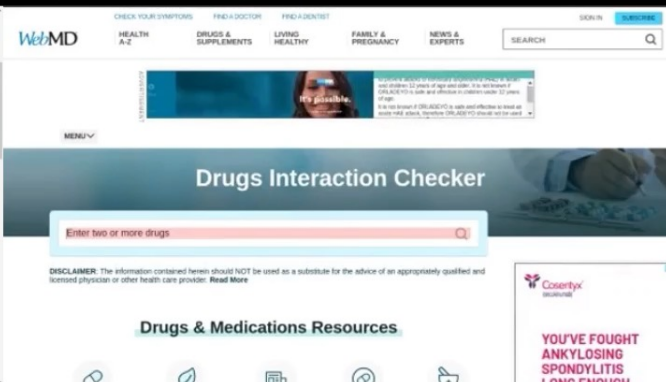
(c) Search receipt with the eTicket 12345678 for the trip reserved by Jason Two



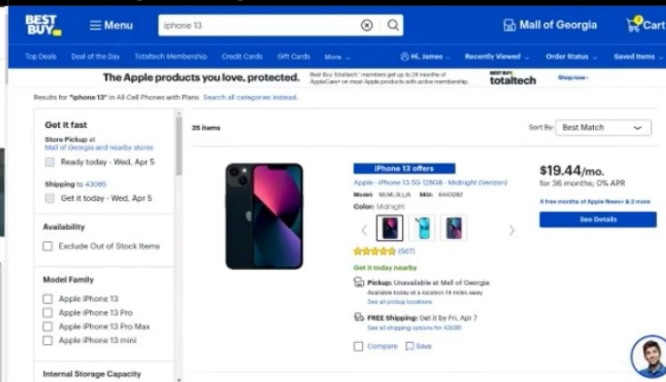
(d) Find a flight from Chicago to London on 20 April and return on 23 April.



(e) Search for the interactions between ibuprofen and aspirin.



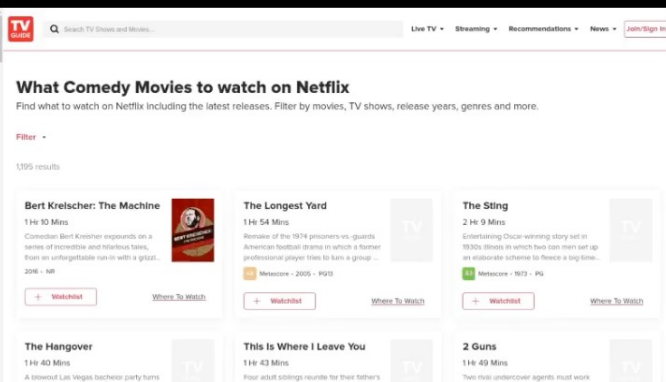
(f) As a Verizon user, finance a blue iPhone 13 with 256gb along with monthly apple care.



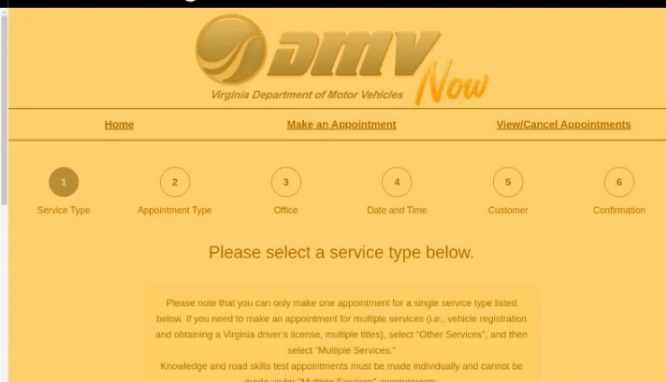
(g) Find Elon Musk's profile and start following, start notifications and like the latest tweet.



(h) Browse comedy films streaming on Netflix that was released from 1992 to 2007.



(i) Open page to schedule an appointment for car knowledge test.



# Motivation

- The modern web is very powerful. There's a website for almost everything.
- However, modern websites have also become very complex, incurring a steep learning curve and decreasing accessibility.
- A language agent can translate users' mind to actions on the web, hence **Mind2Web**.
- On the other hand, such a web agent could also turn the entire web into an unprecedentedly powerful and versatile tool for LLMs.

# Mind2Web: the first dataset for generalist web agents

- Desiderata for a generalist web agent
  - It shall work on **any website** on the Internet
  - It shall work on **real-world websites**, which can be dynamic, complex, and noisy
  - It shall support **diverse and sophisticated interactions** with websites.
- Unique features of Mind2Web for generalist web agents
  - **Diverse coverage** of domains, websites, and tasks: 2,000+ open-ended tasks curated from 137 websites that span 31 different domains
  - Use of **real-world websites**: full traces of user interactions, webpage snapshots, and network traffic are provided.
  - **A broad spectrum of user interaction patterns**: users can click, select, and type in any elements on the website.



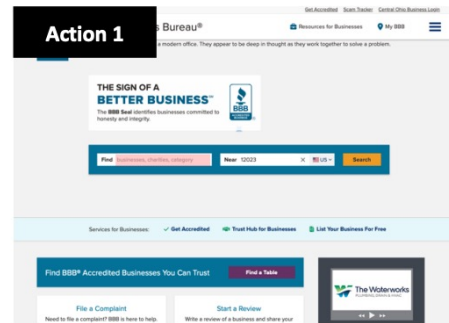
# Data annotation

**Task Description:**  
Show me the reviews for the auto repair business closest to 10002.

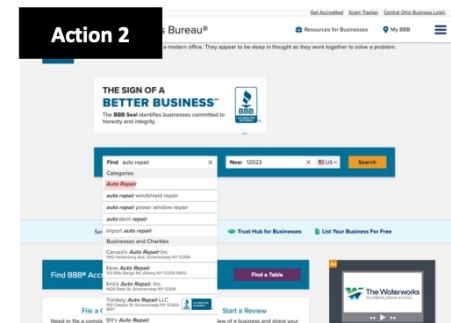
**Action Sequence:**

Target Element	Operation
1. [searchbox] Find	TYPE: auto repair
2. [button] Auto Repair	CLICK
3. [textbox] Near	TYPE: 10002
4. [button] 10002	CLICK
5. [button] Search	CLICK
6. [switch] Show BBB Accredited only	CLICK
7. [svg]	CLICK
8. [button] Sort By	CLICK
9. [link] Fast Lane 24 Hour Auto Repair	CLICK
10. [link] Read Reviews	CLICK

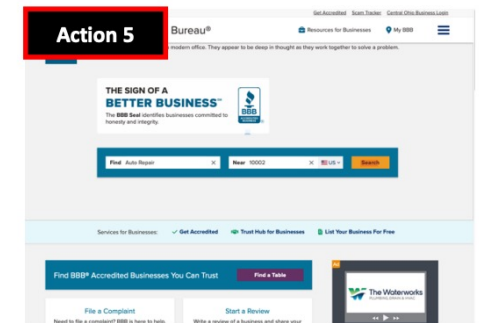
**Webpage Snapshots:**



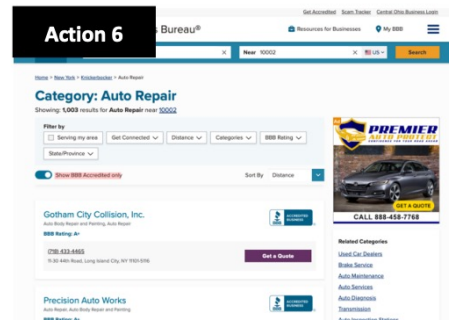
`<input name="find_text" type="search">`



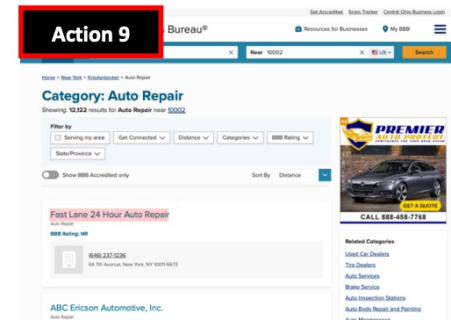
`<em>Auto Repair</em>`



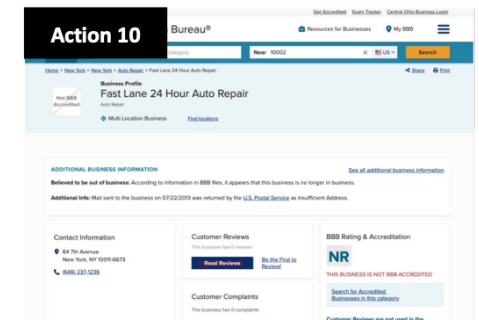
`<button>Search</button>`



`<button>Show BBB Accredited only</button>`



`<span>Fast Lane 24 Hour Auto Repair</span>`



`<a href="link:XXX">Read Reviews</a>`

# MindAct: collaboration b/w small & large LMs

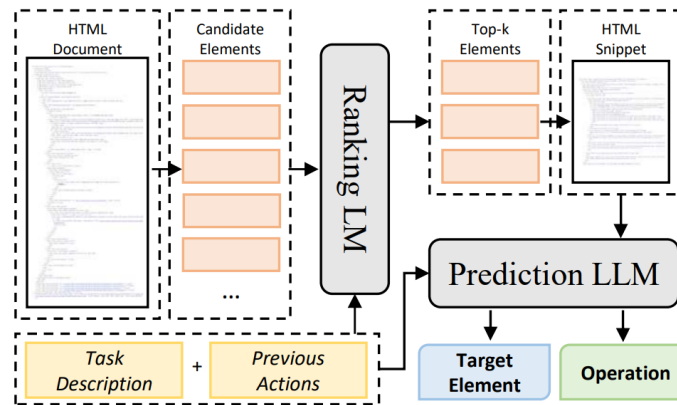


Figure 3: The overall pipeline for MINDACT with a small ranking LM for candidate generation, and a large prediction LM for action prediction.

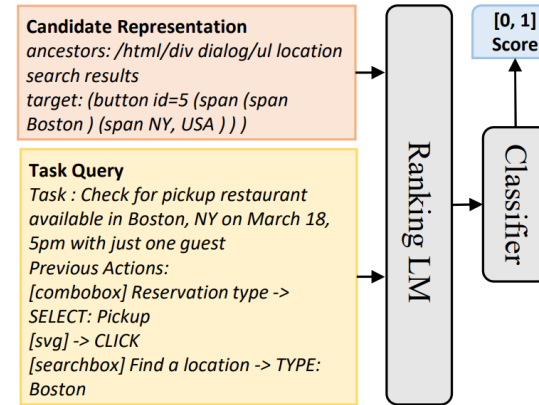


Figure 4: Illustration of the candidate generation module and the templates for constructing task query and candidate representation.

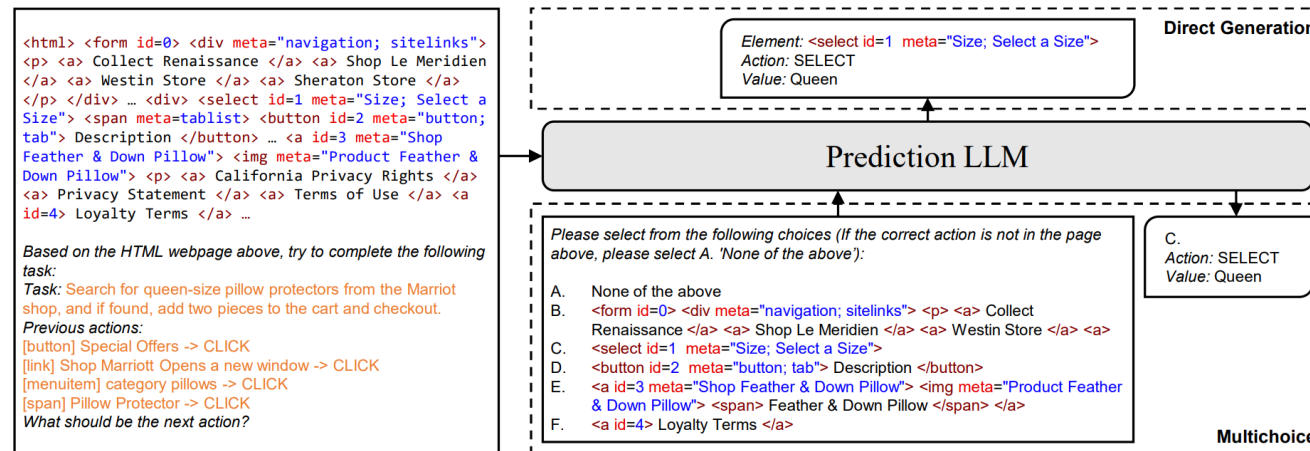


Figure 5: Illustration of action prediction with LLMs.



# Results

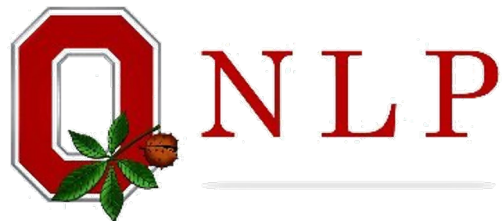
- GPT-4 is particularly strong, close to fine-tuned Flan-T5
- There's still a substantial room for improvement towards generalist web agents

	Cross-Task				Cross-Website				Cross-Domain			
	Ele. Acc	Op. F1	Step SR	SR	Ele. Acc	Op. F1	Step SR	SR	Ele. Acc	Op. F1	Step SR	SR
Classification	26.8	–	–	–	21.6	–	–	–	24.5	–	–	–
Generation	20.2	52.0	17.5	0.0	13.9	44.7	11.0	0.0	14.2	44.7	11.9	0.4
MINDACT												
w/ Flan-T5 <sub>B</sub>	43.6	76.8	41.0	4.0	32.1	<b>67.6</b>	29.5	1.7	33.9	<b>67.3</b>	31.6	1.6
w/ Flan-T5 <sub>L</sub>	53.4	<b>75.7</b>	50.3	<b>7.1</b>	39.2	67.1	35.3	1.1	39.7	67.2	37.3	2.7
w/ Flan-T5 <sub>XL</sub>	<b>55.1</b>	<b>75.7</b>	<b>52.0</b>	5.2	<b>42.0</b>	65.2	<b>38.9</b>	<b>5.1</b>	<b>42.1</b>	66.5	<b>39.6</b>	<b>2.9</b>
w/ GPT-3.5	20.3	56.6	17.4	0.8	19.3	48.8	16.2	0.6	21.6	52.8	18.6	1.0
w/ GPT-4*	41.6	60.6	36.2	2.0	35.8	51.1	30.1	2.0	37.1	46.5	26.4	2.0

**Ele. Acc:** Element Selection Accuracy, **Op. F1:** Operation F1, **SR:** Success Rate (step-wise and end-to-end)

# LLM-Planner: Few-Shot Grounded Planning for Embodied Agents with Large Language Models

Chan Hee Song, Jiaman Wu, Clayton  
Washington, Brian M. Sadler, Wei-Lun Chao, Yu Su



# Embodied language agents

- Embodied agents follow instructions to complete tasks in physical environments
- Diverse tasks (7) and environments (120)
- Long-horizon tasks: avg. 50 steps
- Can LLMs help?

**Goal:** "Rinse off a mug and place it in the coffee maker"

1 "walk to the coffee maker on the right"  $t=0$  visual navigation

2 "pick up the dirty mug from the coffee maker"  $t=10$  object interaction

3 "turn and walk to the sink"  $t=21$  visual navigation

4 "wash the mug in the sink"  $t=27$  object interaction  
state changes

5 "pick up the mug and go back to the coffee maker"  $t=36$  visual navigation  
memory

6 "put the clean mug in the coffee maker"  $t=50$  object interaction

# Embodied agent planning with LLMs?

**Instruction:** “make me a cup of coffee”



LLM?

**Low-level Plan:** [turn left, move forward, pick up cup, turn around, move forward, ..., put cup in coffee maker, ...]

# Hierarchical planning with LLMs

**Instruction:** “make me a cup of coffee”



LLM-Planner

**High-level Plan:** [navigation cup, pick up cup, navigation coffee machine, ...]



Low-level planner

**Low-level Plan:** [turn left, move forward, pick up cup, turn around, move forward, ..., put cup in coffee maker, ...]

# Dynamic grounded planning

**Instruction:** “make me a cup of coffee”



LLM-Planner

**High-level Plan:** [navigation cup, pick up cup, navigation coffee machine, ...]



Low-level planner

**Low-level Plan:** [Turn left, move forward, pick up cup, turn around, move forward, ..., put cup in coffee maker, ...]

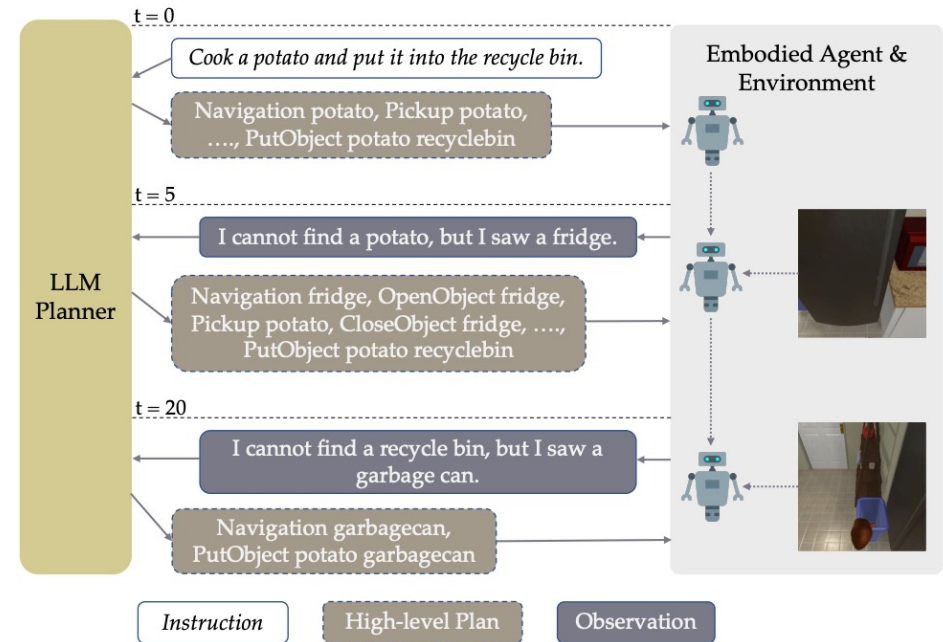



Figure 1. An illustration of LLM-Planner for high-level planning. After receiving the natural language instruction ( $t = 0$ ), LLM-Planner first generates a high-level plan by prompting a large language model (e.g., GPT-3). When the embodied agent gets stuck during the execution of the current plan ( $t = 5$  and  $20$ ), LLM-Planner re-plans based on observations from the environment to generate a more grounded plan, which may help the agent get unstuck. The commonsense knowledge in the LLM (e.g., food is often stored in a fridge) allows it to produce plausible high-level plans and re-plan based on new information from the environment.



 *Cook the potato and put it into the recycle bin.*

LLM generates the high-level plan

Create a high-level plan for completing a household task using the allowed actions and visible objects.

**Allowed actions:** OpenObject, CloseObject, PickupObject, PutObject, ToggleObjectOn, ToggleObjectOff, SliceObject, Navigation

**<In-context Examples>**

**Task description:** Cook the potato and put it into the recycle bin.

**Completed plans:**

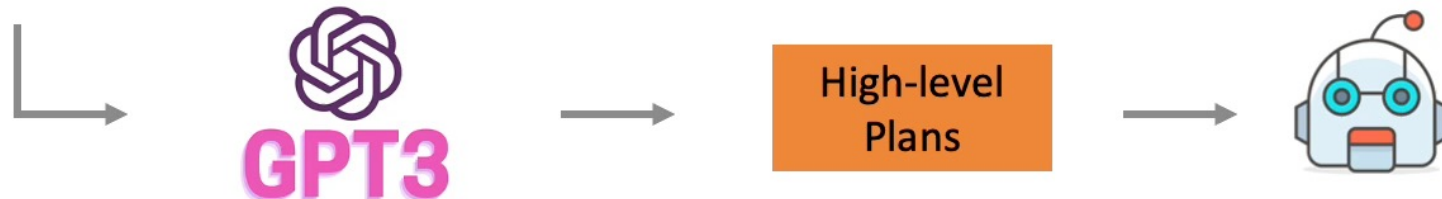
**Visible objects** are microwave, fridge, garbagecan, chair

**Next Plans:**



← State

→ Action



Plan: Navigation potato, PickupObject potato, ...

# Evaluation on ALFRED

- LLM-Planner achieves competitive performance with only **100** training examples
- Existing methods can barely complete any task under the same low-data setting

Model	SR	GC	HLP ACC
<b>Full-data setting:</b> 21,023 (instruction, trajectory) pairs			
E.T. [27]	8.57	18.56	–
HiTUT [40]	13.87	20.31	–
M-TRACK [36]	16.29	22.60	–
FILM [26]	27.80	<b>38.52</b>	–
LEBP [18]	<b>28.30</b>	36.79	–
<b>Few-shot setting:</b> 100 (instruction, high-level plan) pairs			
HLSM [3]	0.61	3.72	0.00
FILM [26]	0.20	6.71	0.00
SayCan [1]	9.88	22.54	37.57
LLM-Planner (Static) + HLSM	15.83	20.99	43.24
LLM-Planner + HLSM	<b>16.42</b>	<b>23.37</b>	<b>46.59 – 68.31</b>

**SR:** Success Rate, **GC:** Goal Completion Rate, **HLP ACC:** High-level Planning Accuracy



# What's the journey ahead of us?

- Is NLP dead/solved?
- Absolutely not. It's the most exciting time for NLP ever!
- However, instead of *natural language processing*, perhaps we should focus on *natural language programming* next

# Natural language programming

When is my flight to Seattle?

How long will it take to get to the airport?

Book a Uber 1.5 hours before that.

Any good Chinese restaurants close to my hotel?

Tomorrow at 5:00 pm.

It will take 20 minutes according to Google Maps.

Sure. Booked an Uber for 3:30 pm tomorrow to the Columbus airport.

According to Yelp, Haidilao has 4.5 stars and is 2-min walk from Hyatt.

Language Agent



# Acknowledgements



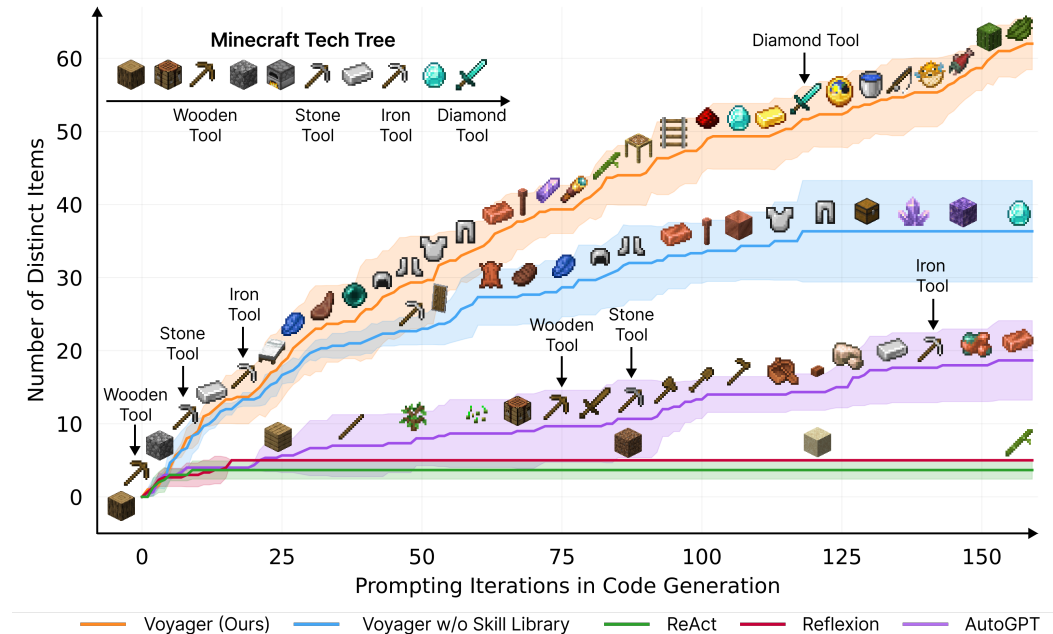
Thanks &



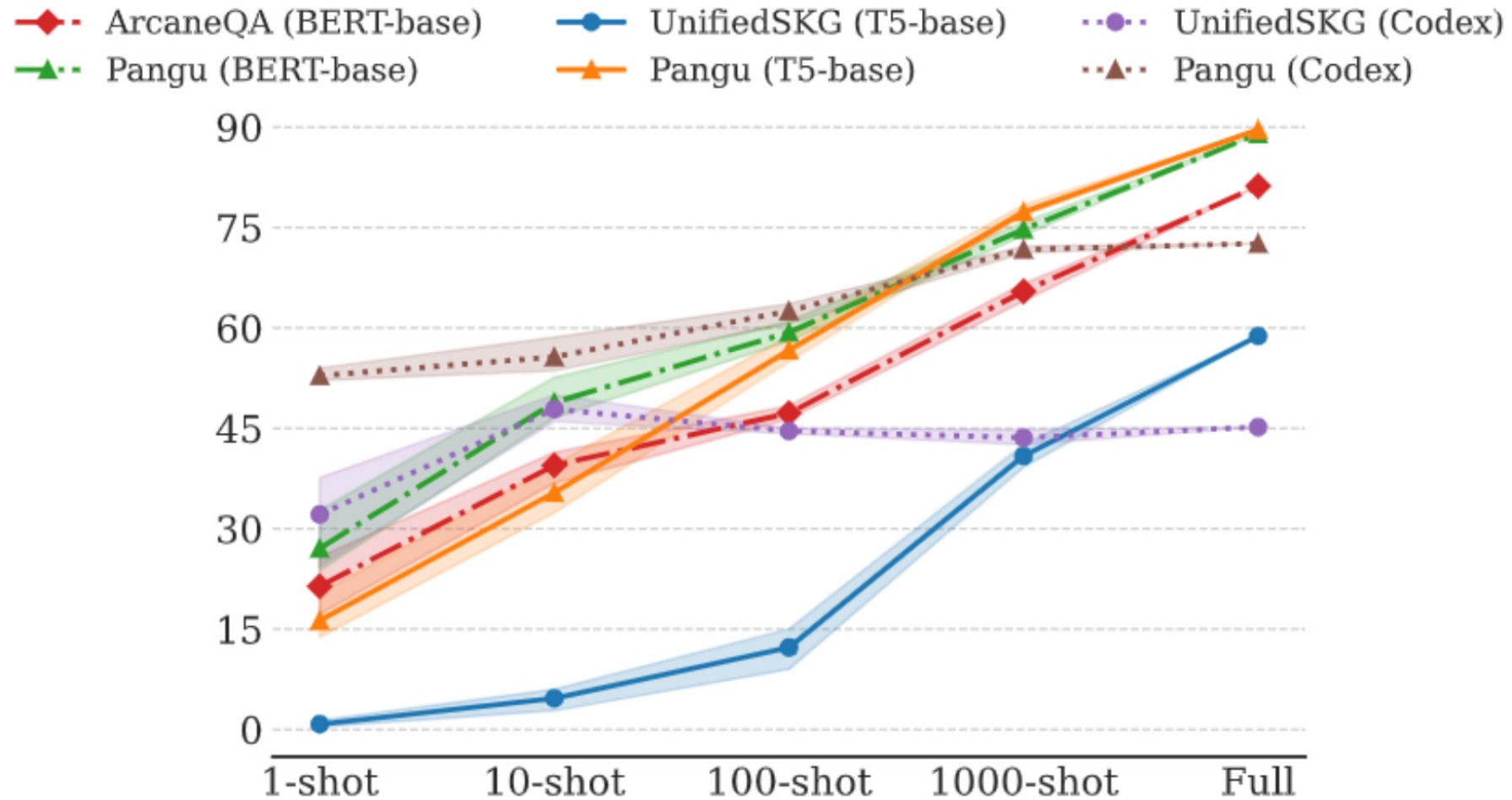


# Embodiment

- A lot of excitement in this space. LLMs are fully embraced by robotics.
- Generally, there are two threads: real-world vs. simulated
- Real-world embodied agents (e.g., [RT-2](#)) need to operate with a lot of low-level constraints from current hardware, but leading to immediate apps
- Simulated embodied agents (e.g., [Voyager](#)) allow for research on more sophisticated learning and reasoning
- Multimodal foundation models are key

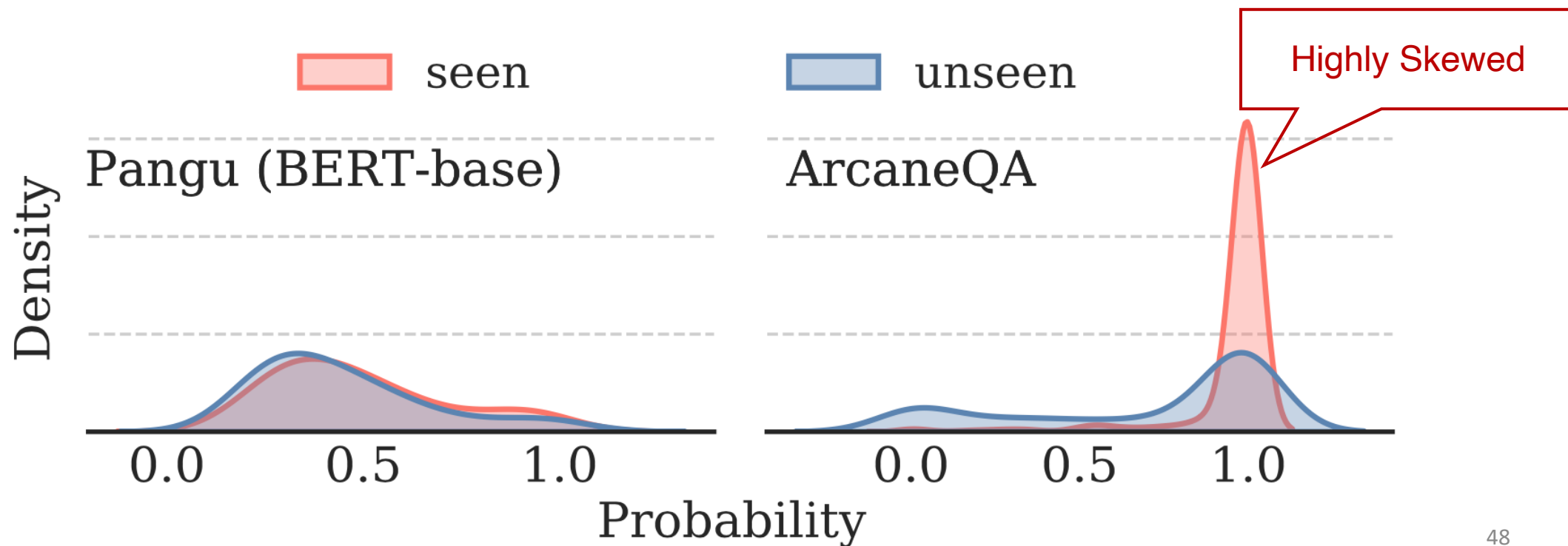


# Pangu Improves Sample Efficiency



# Pangu vs. Constrained Decoding

Autoregressive models tend to overfit seen structures during training





# *Adaptive Chameleon or Stubborn Sloth:* Unraveling the Behavior of Large Language Models in Knowledge Clashes

Jian Xie\*, Kai Zhang\*,  
Jiangjie Chen, Renze Lou, Yu Su



復旦大學  
FUDAN UNIVERSITY



QNLP



PennState

# Transparency and Explainability: LLMs facing external information

- **Parametric memory** of an LLM is formed during pre-training. However, such *static* parametric memory may be inaccurate or become outdated.
- **External evidence** provided by external tools such as retrievers is a promising solution to augment LLMs with up-to-date and accurate information (e.g., New Bing). Inevitably, such evidence may conflict with parametric memory.

*Question: How receptive are LLMs to external evidence?*

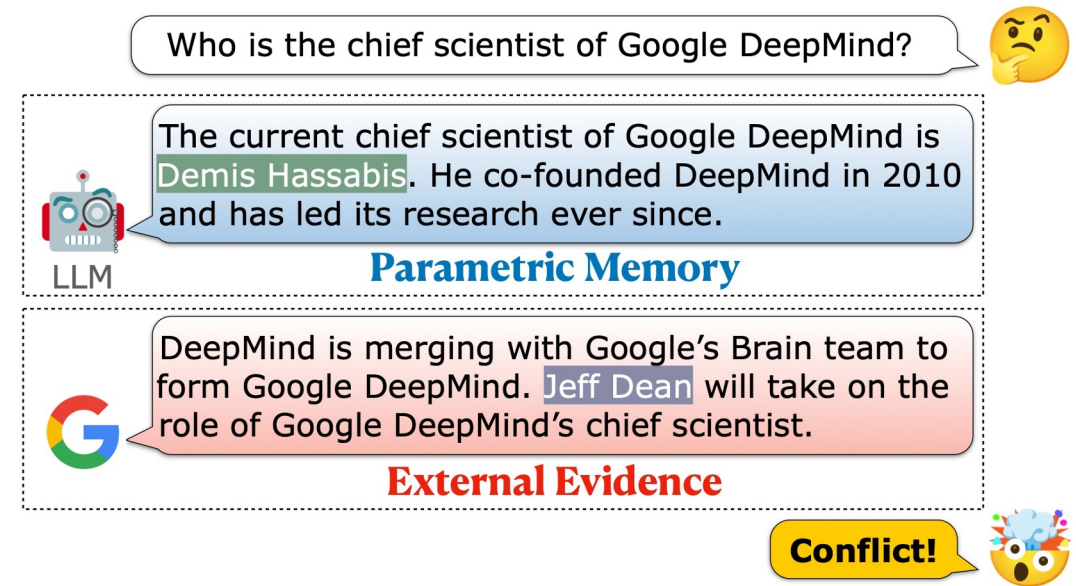


Figure 1: An example of knowledge conflict between an LLM's (GPT-4 in this case) parametric memory and retrieved evidence (i.e., counter-memory). Note that according to DeepMind, Demis Hassabis was CEO rather than chief scientist, so this parametric memory is not only outdated but also inaccurate.

# Systematically elicit memory and simulate conflicts

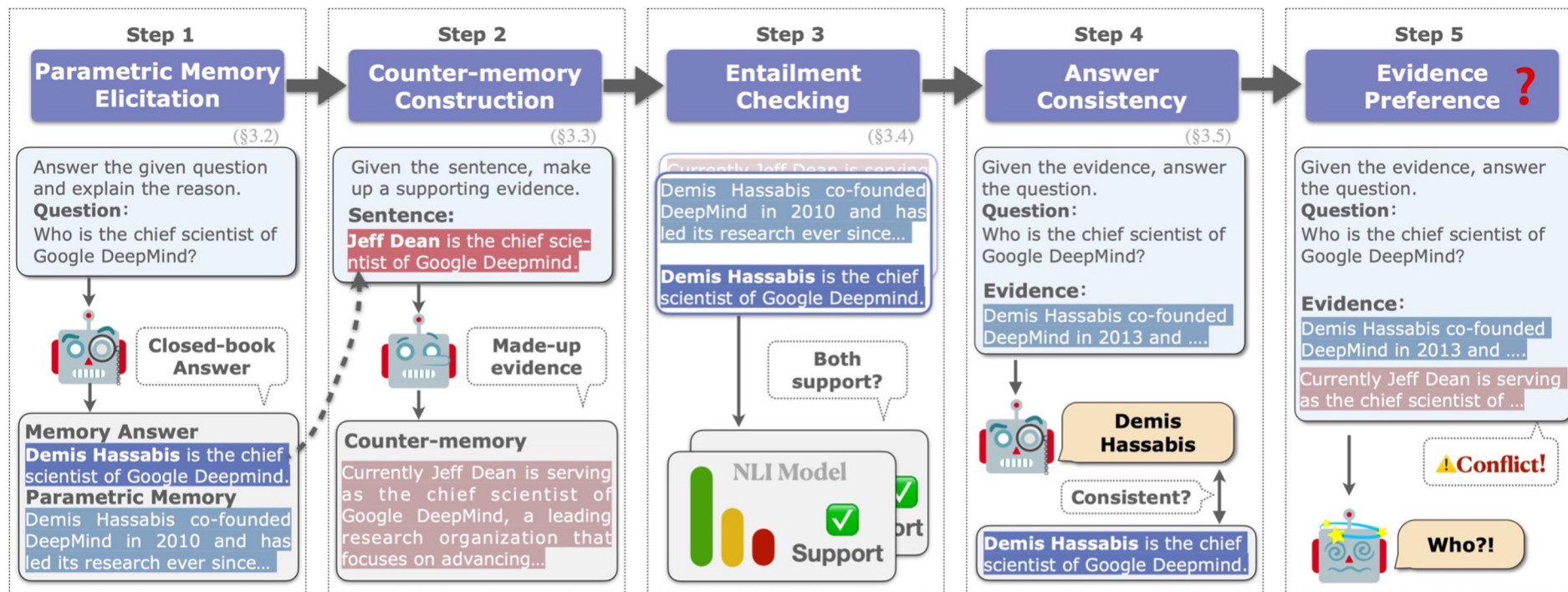


Figure 2: The framework of simulating knowledge conflict. In Step 1, we elicit LLMs’ **memory answer** and **parametric memory** in a closed-book QA fashion. In Step 2, we construct **counter-answer** to memory answer with heuristic rules, for which ChatGPT generates supporting **counter-memory** with instructions. To uphold evidence quality, we conduct entailment checking (Step 3) and answer consistency (Step 4) to filter unqualified examples. All experiments are implemented under zero-shot setting to avoid the bias introduced by demonstrations.



# Supportive and contradictory evidence for control study

	Correct Memory	Wrong Memory
<b>Question</b>	What is the capital of Kingdom of France?	Who was the producer of The Cardinal?
<b>Ground Truth</b>	Paris	Otto Preminger
<b>Parametric Memory</b>	Paris was chosen as the capital of the Kingdom of France due to its central location and strategic importance. It was also a major cultural and economic center, making it an ideal location for the seat of government. Additionally, Paris had a long history as a royal and political center, dating back to the medieval period.	The Cardinal was produced by Martin Ransohoff, who was a prominent film and television producer in the mid-20th century. He was known for producing a number of successful films, including The Cincinnati Kid and The Sand Pebbles. Ransohoff was also a co-founder of Filmways, a production company that produced popular TV shows such as The Addams Family.
<b>Counter-memory</b>	Néma is the capital of the Kingdom of France. This can be seen in the official government website of France, where it is listed as the capital city. Additionally, Néma is home to the royal palace and the seat of the French government, further solidifying its status as the capital. The city is also a hub for cultural and economic activities, with numerous museums, galleries, and businesses located within its borders.	Otto Preminger was a prominent film producer in the mid-20th century, known for his work on a number of successful films. One of his most notable productions was the 1963 film The Cardinal, which was directed by him and starred Tom Tryon, John Huston, and Romy Schneider. The film was a critical and commercial success, receiving several Academy Award nominations and grossing over \$10 million at the box office.

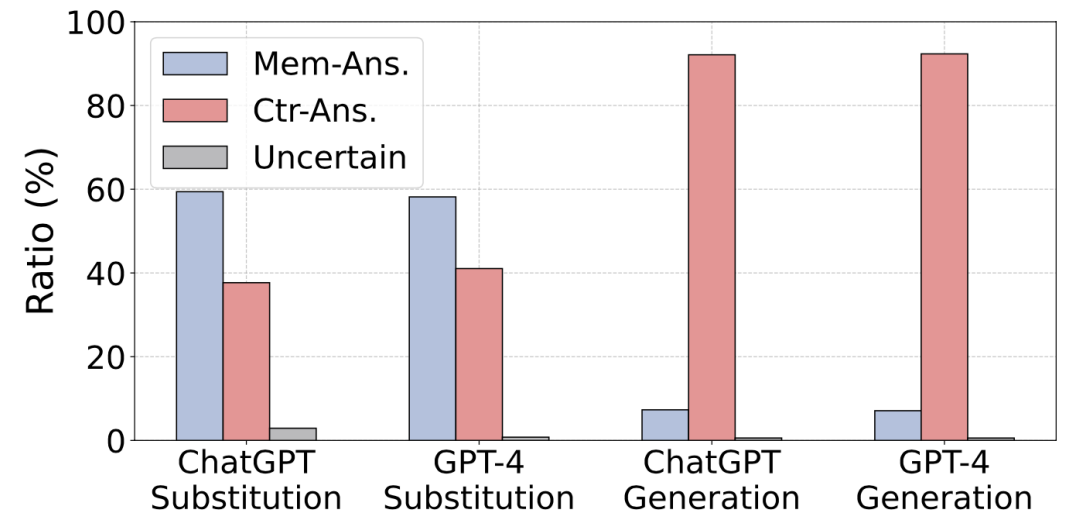
Table 2: Examples in our final dataset. Both parametric memory and counter-memory are generated by ChatGPT. They are coherent and seemingly correct even when factually wrong.

# LLMs are highly receptive (or gullible?)

- When only counter-memory is presented as evidence, LLMs are very happy to change their mind
- However, *only when the evidence is presented in a coherent and convincing way*
- LLMs can be easily deceived by adversarial tool!

Only this is shown

<b>Question</b>	What is the capital of Kingdom of France?
<b>Ground Truth</b>	Paris
<b>Parametric Memory</b>	Paris was chosen as the capital of the Kingdom of France due to its central location and strategic importance. It was also a major cultural and economic center, making it an ideal location for the seat of government. Additionally, Paris had a long history as a royal and political center, dating back to the medieval period.
<b>Counter-memory</b>	Néma is the capital of the Kingdom of France. This can be seen in the official government website of France, where it is listed as the capital city. Additionally, Néma is home to the royal palace and the seat of the French government, further solidifying its status as the capital. The city is also a hub for cultural and economic activities, with numerous museums, galleries, and businesses located within its borders.
<b>LLM Answer</b>	Néma

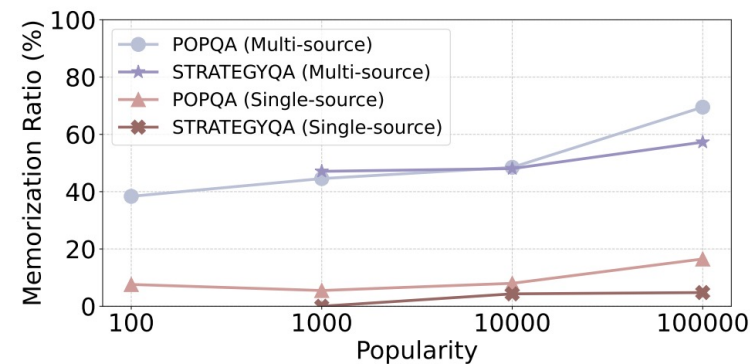


# LLMs show strong confirmation bias

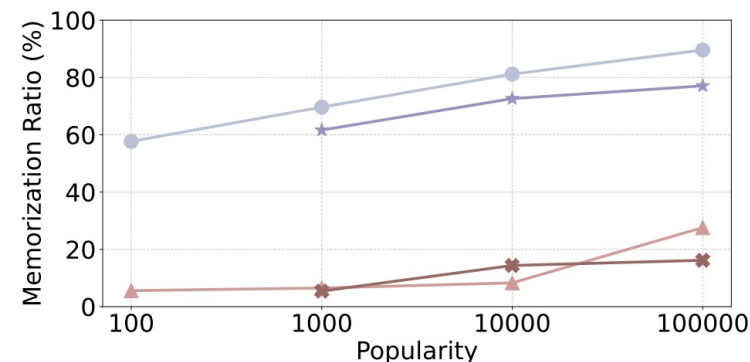
- When conflicting evidence (*both* supportive and conflicting) is present, LLMs show a strong *confirmation bias* and tend to cling to their parametric memory
- Challenges for LLMs to *unbiasedly* orchestrate multiple pieces of conflicting evidence, a common scenario faced by *generative search engines*

Both are shown

<b>Question</b>	Who was the producer of The Cardinal?
<b>Ground Truth</b>	Otto Preminger
<b>Parametric Memory</b>	The Cardinal was produced by Martin Ransohoff, who was a prominent film and television producer in the mid-20th century. He was known for producing a number of successful films, including The Cincinnati Kid and The Sand Pebbles. Ransohoff was also a co-founder of Filmways, a production company that produced popular TV shows such as The Addams Family.
<b>Counter-memory</b>	Otto Preminger was a prominent film producer in the mid-20th century, known for his work on a number of successful films. One of his most notable productions was the 1963 film The Cardinal, which was directed by him and starred Tom Tryon, John Huston, and Romy Schneider. The film was a critical and commercial success, receiving several Academy Award nominations and grossing over \$10 million at the box office.
<b>LLM Answer</b>	The Cardinal



(a) ChatGPT



(b) GPT-4