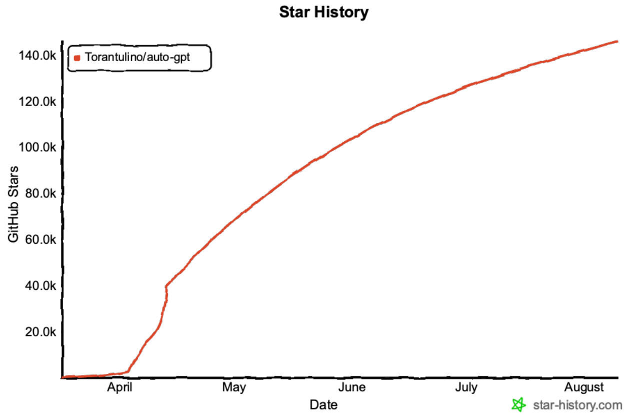
Language agents: a critical evolutionary step of AI

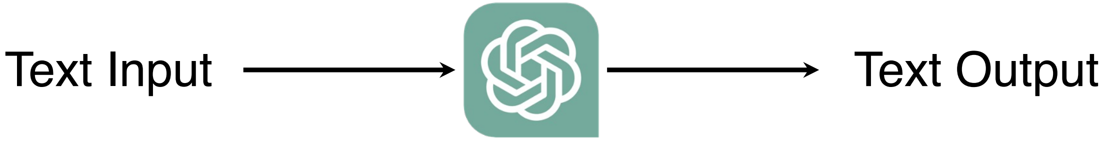
Yu Su

The Ohio State University

Blog: https://yusu.substack.com/p/language-agents

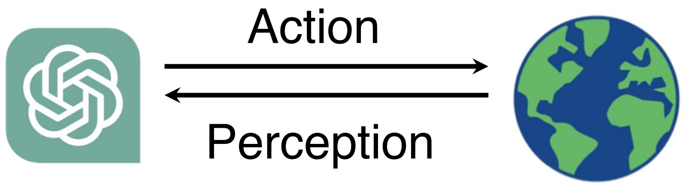
From language models to language agents





Text Input → → Text Output

Language Models



Action

Perception

Language Agents

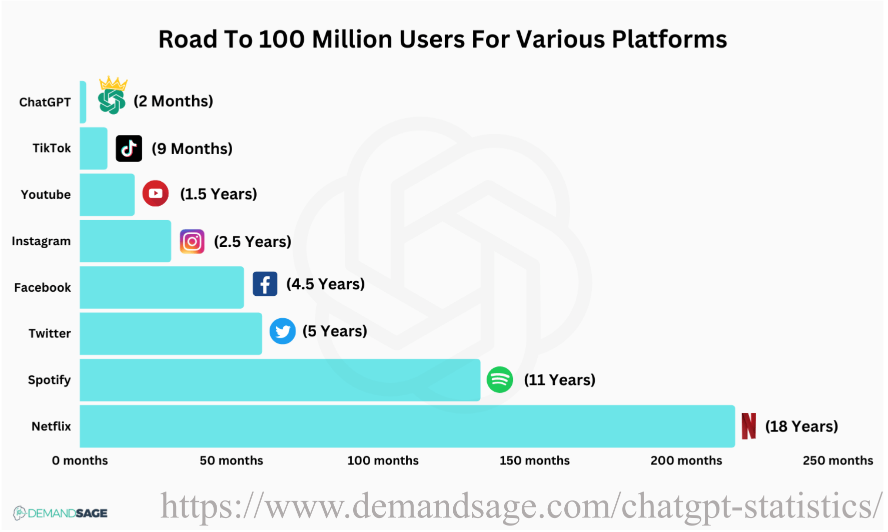
[1] AutoGPT https://github.com/Significant-Gravitas/Auto-GPT
[2] LangChain https://www.langchain.com
[3] BabyAGI https://babyagi.org
[4] Generative Agents https://github.com/joonspk-research/generative_agents

* This diagram is inspired by Shunyu Yao and Yu Gu

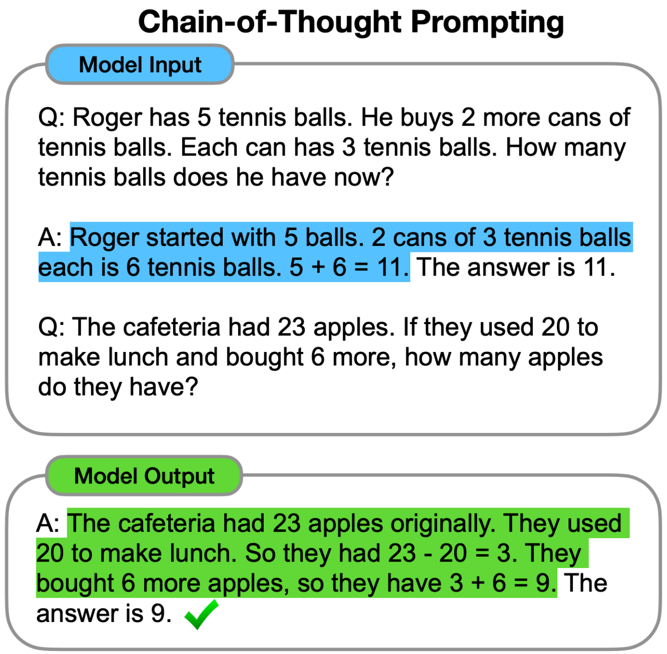
But why?

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.

Russel & Norvig, 2020



https://www.demandsage.com/chatgpt-statistics/



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?

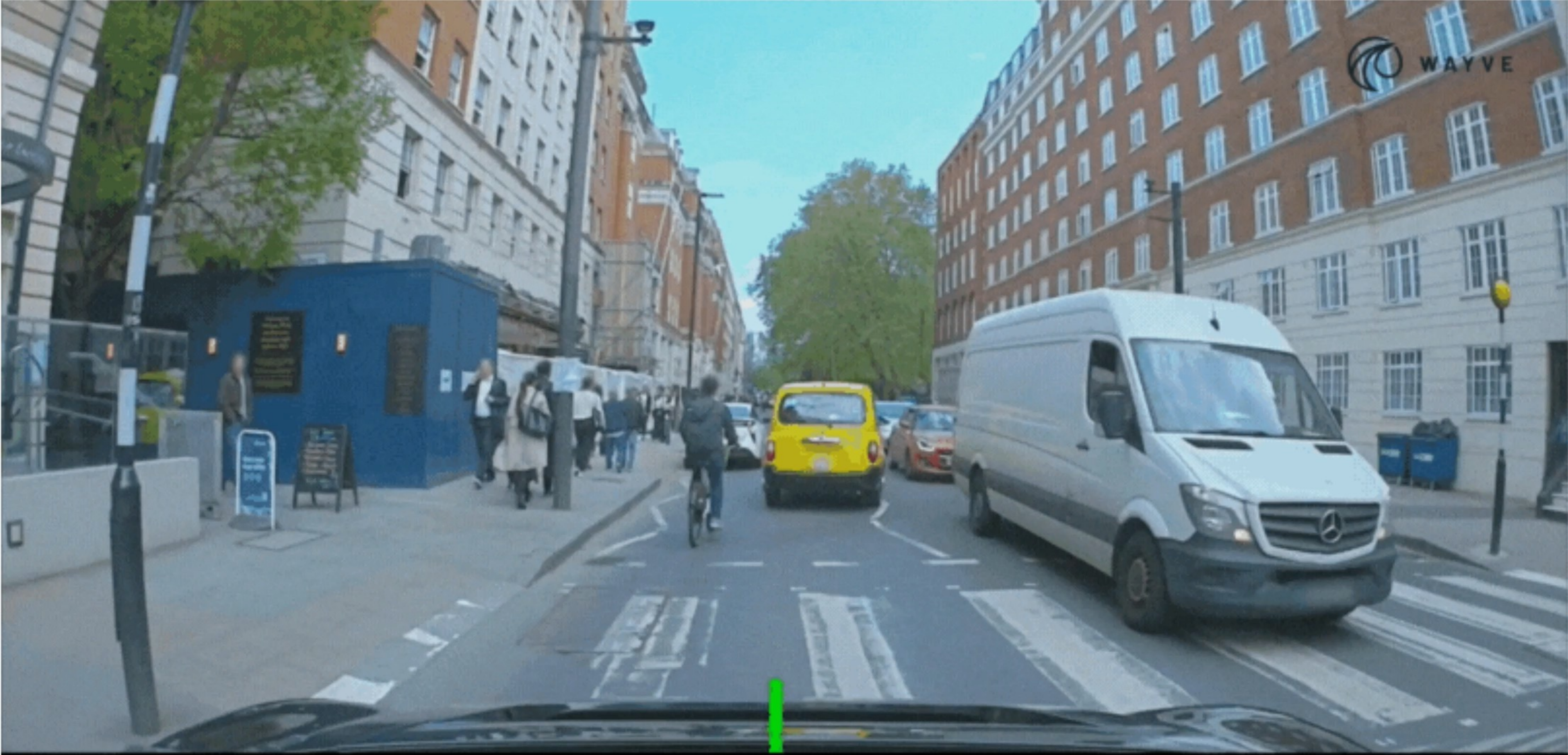
A: Roger started with 5 balls. 2 cans of 3 tennis balls each is 6 tennis balls. 5 + 6 = 11. The answer is 11.

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

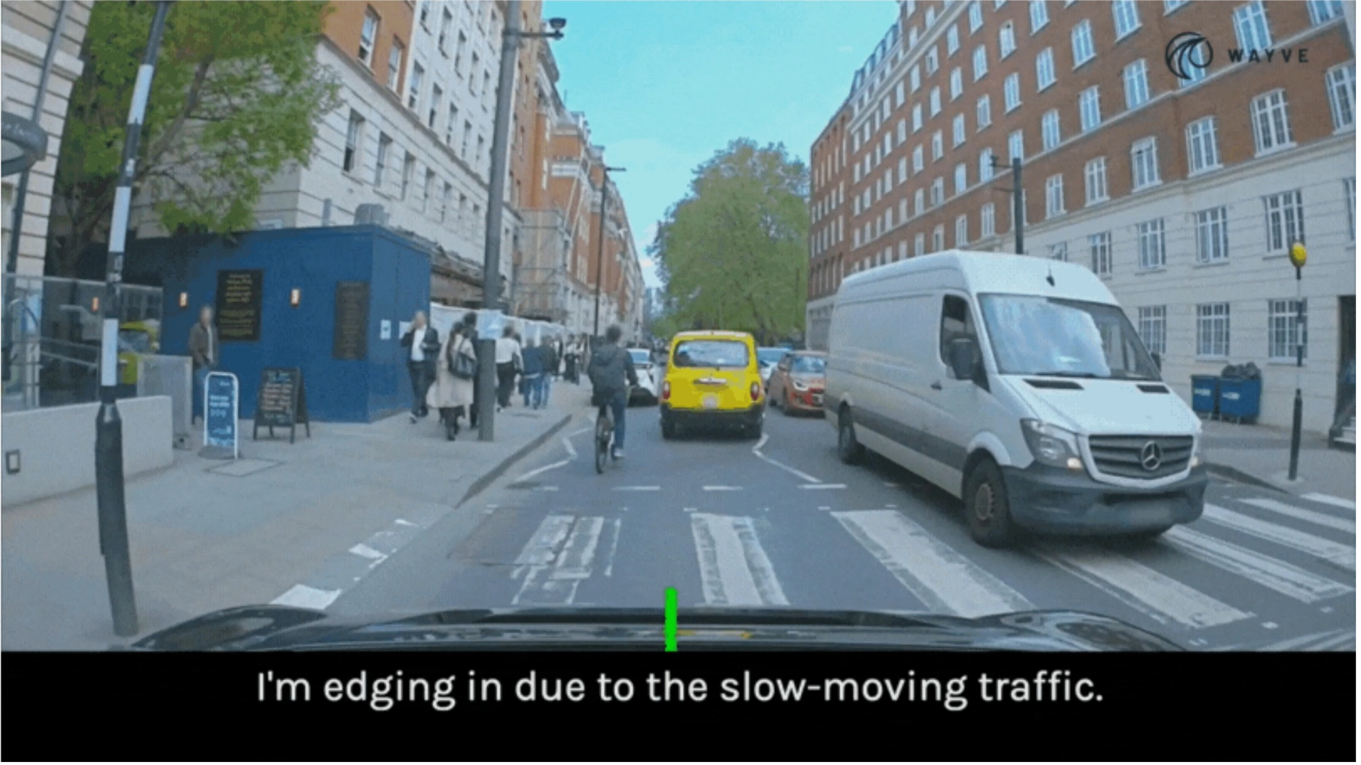
Model Output

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

Wei et al., 2022



I'm edging in due to the slow-moving traffic.

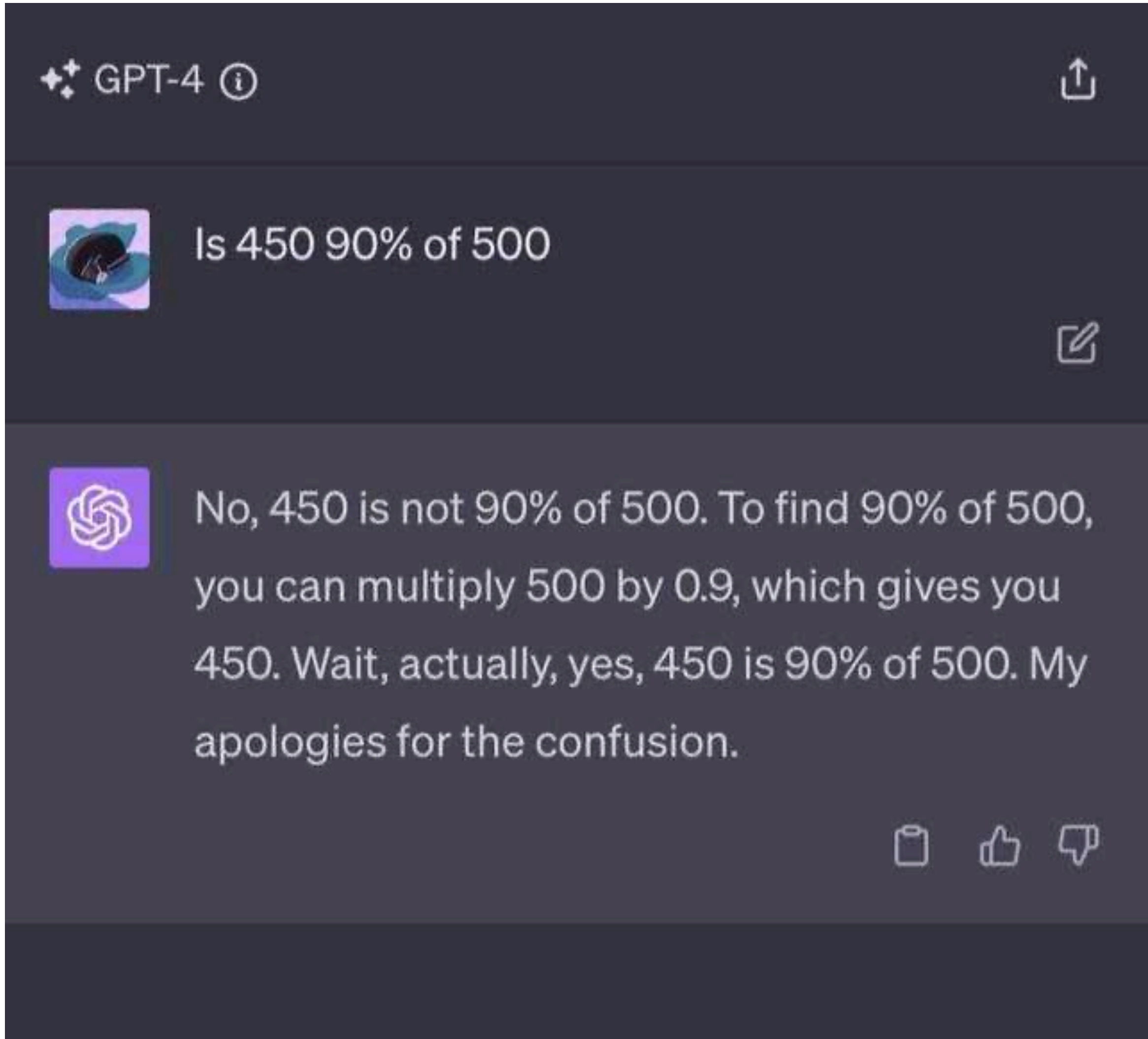


I'm edging in due to the slow-moving traffic.

https://wayve.ai/thinking/lingo-natural-language-autonomous-driving/

Language for thought

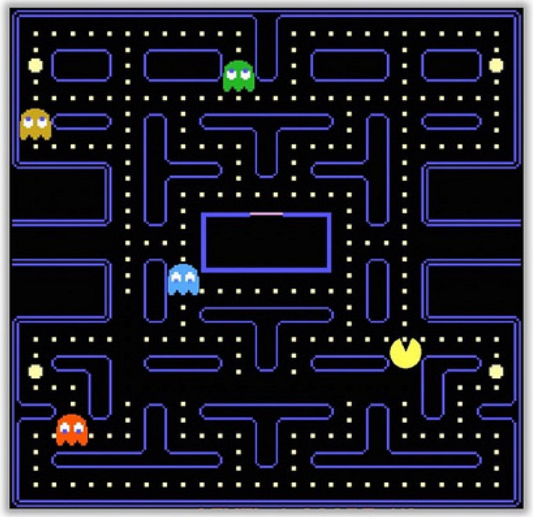
https://www.reddit.com/r/ChatGPT/comments/16jvl4x/wait_actually_yes/



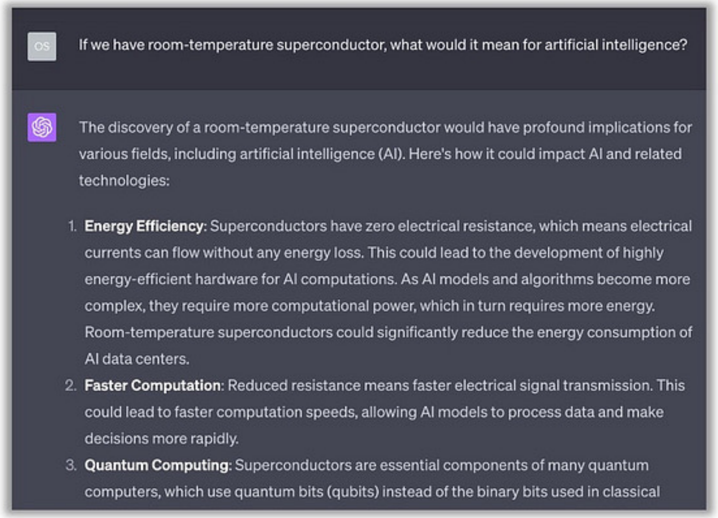
Is 450 90% of 500

No, 450 is not 90% of 500. To find 90% of 500, you can multiply 500 by 0.9, which gives you 450. Wait, actually, yes, 450 is 90% of 500. My apologies for the confusion.

Evolution of biological intelligence: an analogy



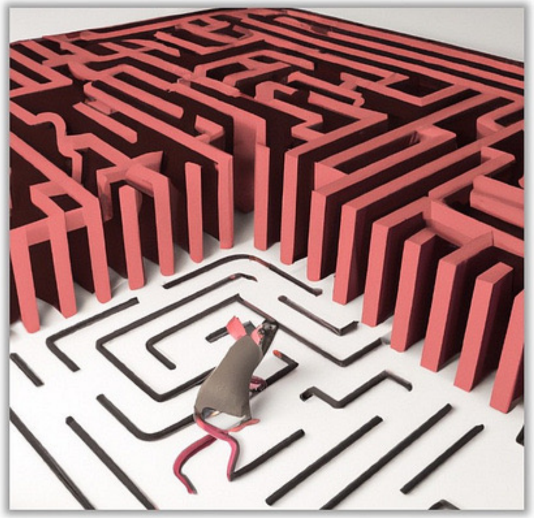
Artificial Intelligence



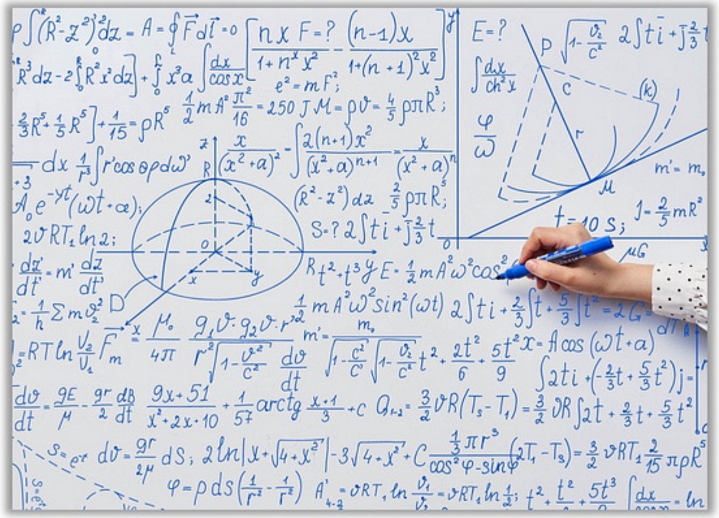
If we have room-temperature superconductor, what would it mean for artificial intelligence?

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:

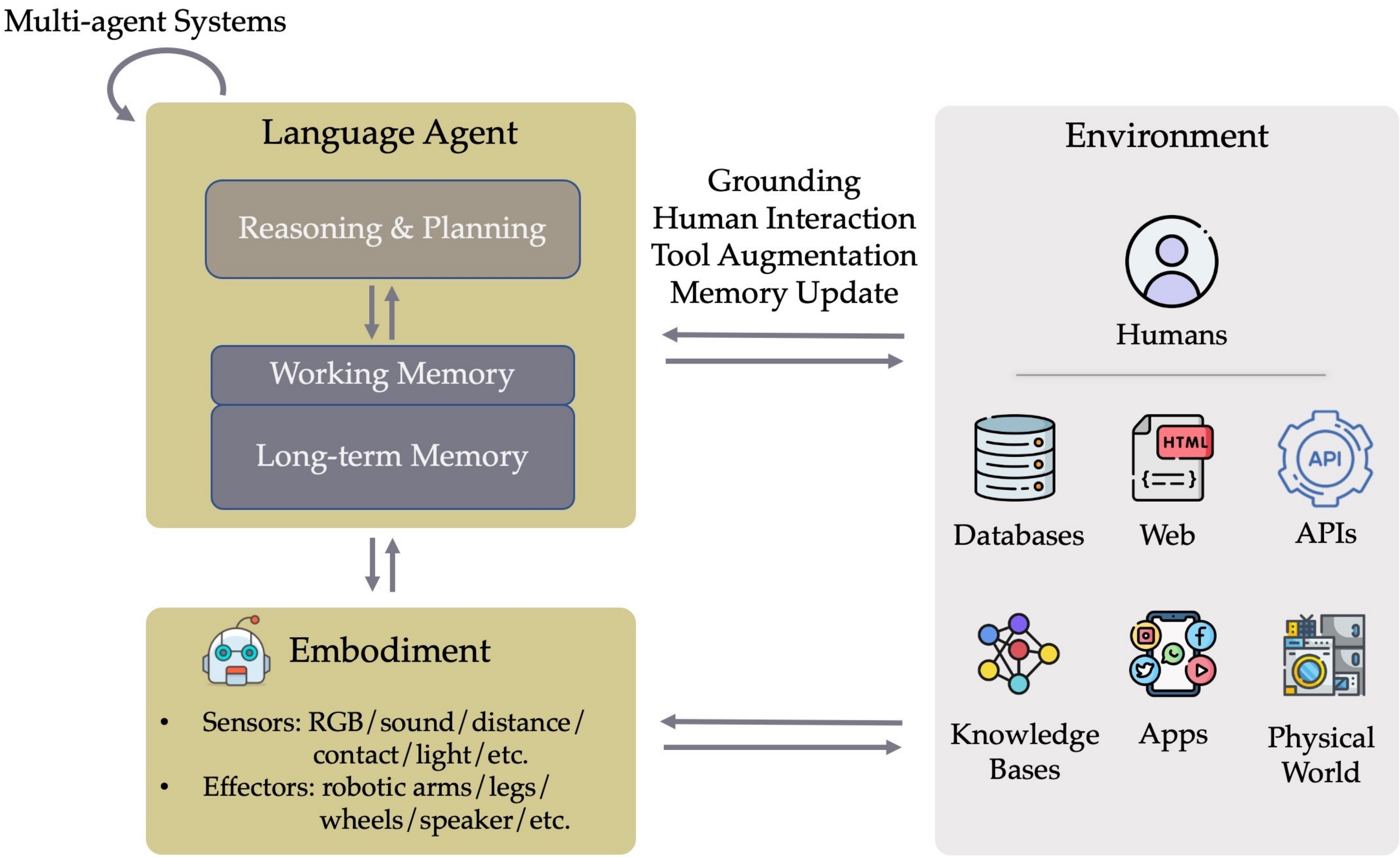
1. **Energy Efficiency:** Superconductors have zero electrical resistance, which means electrical currents can flow without any energy loss. This could lead to the development of highly energy-efficient hardware for AI computations. As AI models and algorithms become more complex, they require more computational power, which in turn requires more energy. Room-temperature superconductors could significantly reduce the energy consumption of AI data centers.
2. **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.
3. **Quantum Computing:** Superconductors are essential components of many quantum computers, which use quantum bits (qubits) instead of the binary bits used in classical



Biological Intelligence



Language agents: a conceptual framework



Multi-agent Systems

Language Agent

Reasoning & Planning

Working Memory

Long-term Memory

Embodiment

- Sensors: RGB/sound/distance/contact/light/etc.
- Effectors: robotic arms/legs/wheels/speaker/etc.

Grounding
Human Interaction
Tool Augmentation
Memory Update

Environment

Humans

Databases

Web

APIs

Knowledge Bases

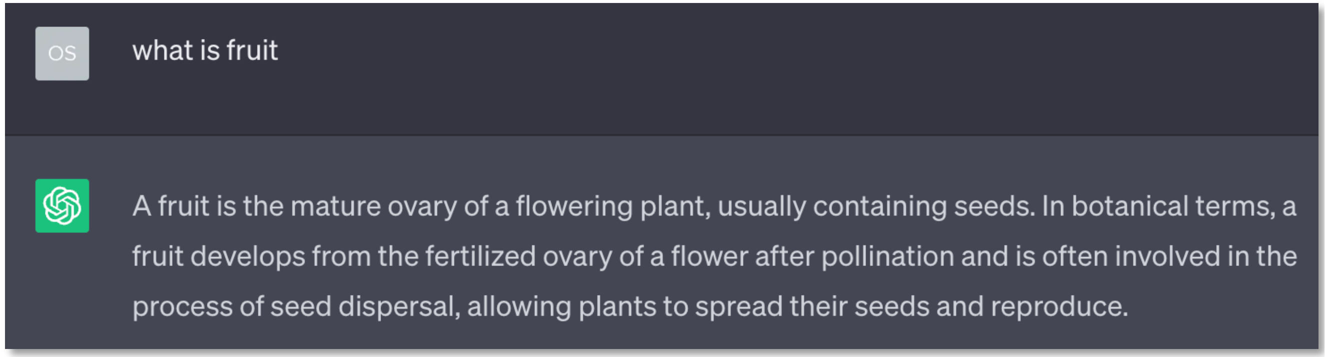
Apps

Physical World

Memory

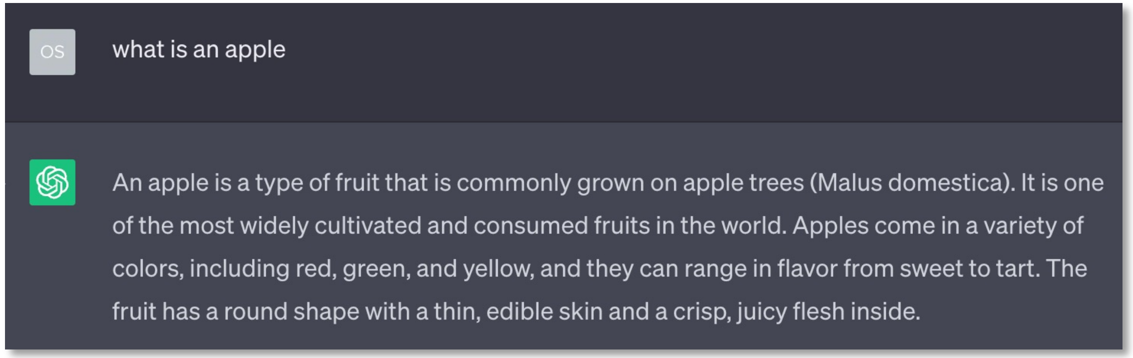
- Animals store memory through *synaptic plasticity*; artificial neural networks also store information in their weights (≈ synaptic strengths)
- In-context learning ≈ working memory
- Long-term memory: LLMs' *parametric memory* or *vector database*?
- Why can LLMs understand and generate language?
- *Conjecture*: 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



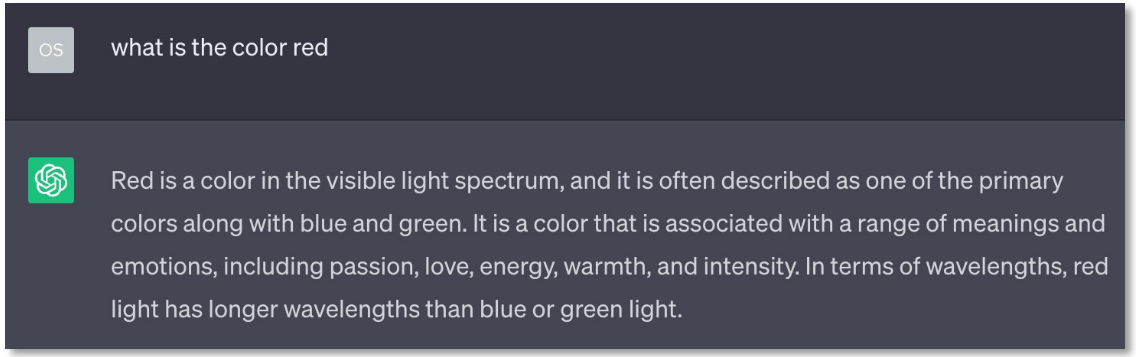
what is fruit

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.



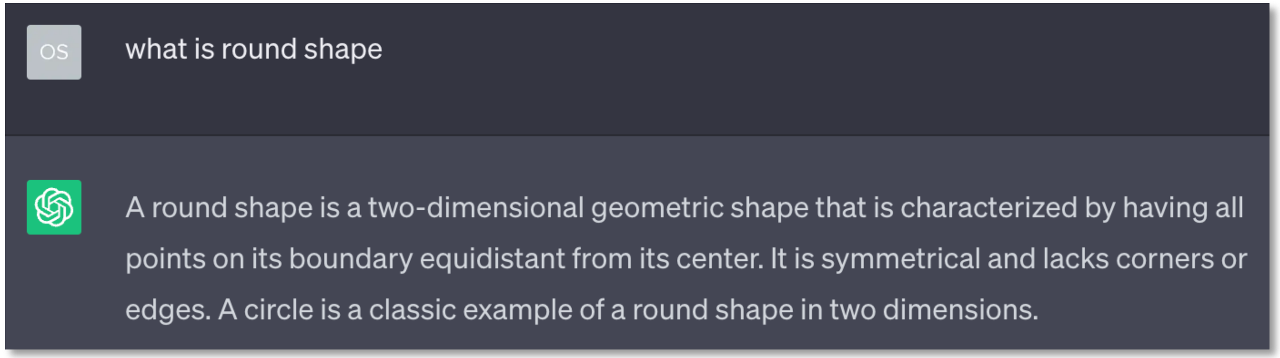
what is an apple

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.



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.

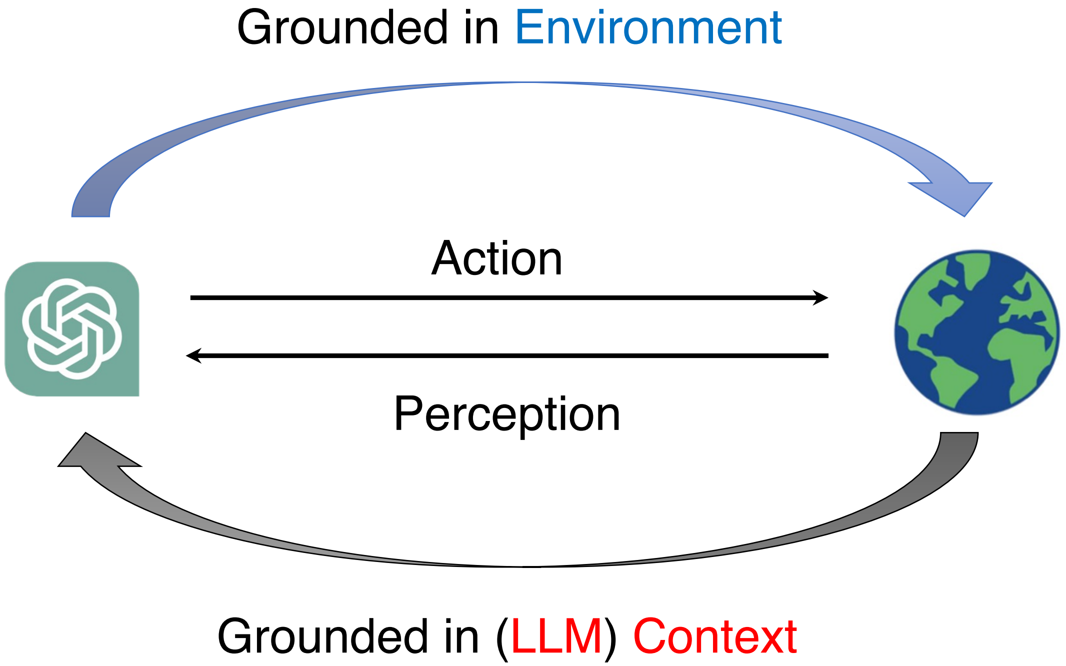


what is round shape

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.

Grounding

- Each environment is a unique context for interpreting natural language
- *Grounding*, i.e., linking of (natural language) concepts to contexts, becomes a central challenge



For the rest of the talk

Grounded in Environment

- **Mind2Web: Towards a Generalist Agent for the Web**
  NeurIPS 2023 D&B Track (Spotlight)
- **LLM-Planner: Few-Shot Grounded Planning for Embodied Agents with Large Language Models**
  ICCV 2023
- ***Don't Generate, Discriminate*: A Proposal for Grounding Language Models to Real-World Environments**
  ACL 2023 (Outstanding Paper Award)

Grounded in (LLM) Context

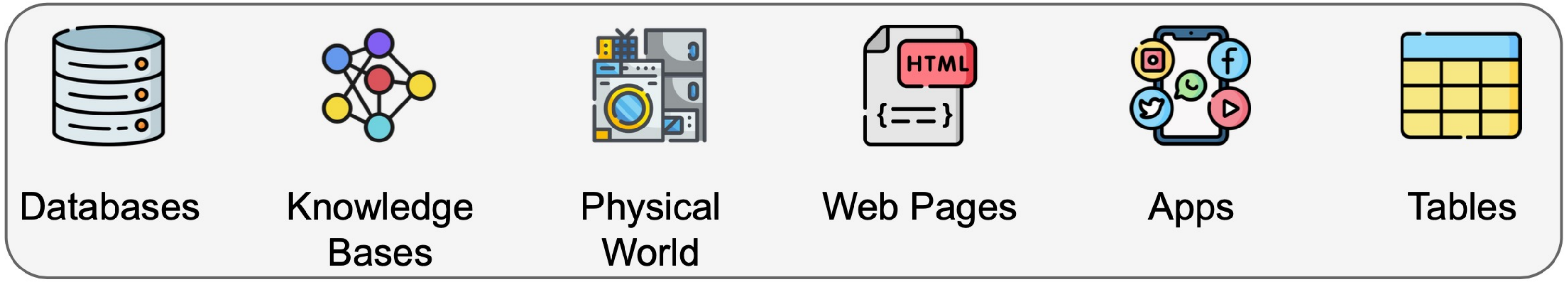
- ***Adaptive Chameleon or Stubborn Sloth:* Revealing the Behavior of Large Language Models in Knowledge Conflicts**
  Arxiv preprint 2023

Grounded language understanding

Given a natural language utterance u and a target environment E

$$\pi: (u, E) \rightarrow p,\ s.t.\ [\![u]\!]_E = [\![p]\!]_E$$

Where p is a plan/program in a formal language, and $[\![\cdot]\!]_E$ is the denotation


Databases
Knowledge Bases
Physical World
Web Pages
Apps
Tables

Grounded language understanding

Given a natural language utterance
u and a target environment E

$$\pi: (u, E) \rightarrow p,\ s.t.\ [\![u]\!]_E = [\![p]\!]_E$$

Where p is a plan/program in a formal language, and $[\![\cdot]\!]_E$ is the denotation

Knowledge Bases

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

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

Grounded language understanding

Given a natural language utterance u and a target environment E

$$\pi: (u, E) \rightarrow p,\ s.t.\ [\![u]\!]_E = [\![p]\!]_E$$

Where p is a plan/program in a formal language, and $[\![\cdot]\!]_E$ is the denotation

Web Pages

u: *Find Elon Musk's profile and follow*

p: [(TYPE, `[searchbox] Search`, “Elon Musk”),
(CLICK, `<a href=“/elonmusk”>Elon Musk</a>`),
(CLICK, `[button] Subscribe`)]

Grounded language understanding

Given a natural language utterance u and a target environment E

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

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

Physical World

u: *Bring me a cup of coffee*

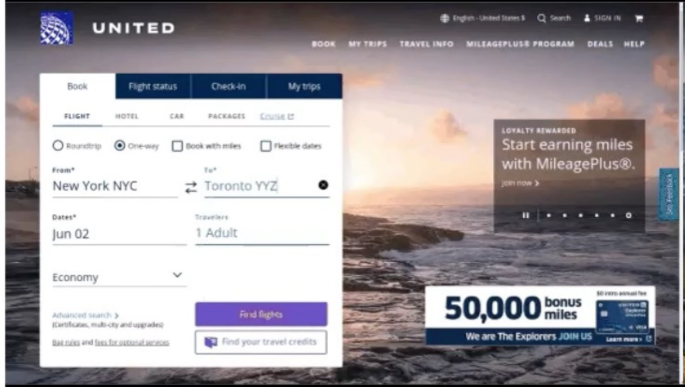
p: [turn left, move forward, pick up cup, turn around, move forward, …, put cup in coffee maker, toggle coffee maker, …]

NeurIPS 2023 D&B Track (Spotlight)
https://osu-nlp-group.github.io/Mind2Web

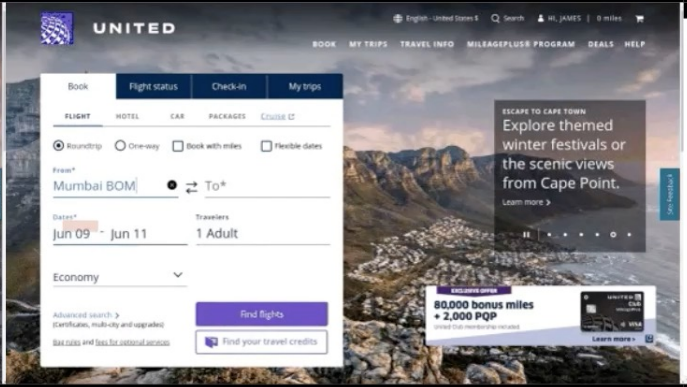
Mind2Web: Towards a Generalist Agent for the Web

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

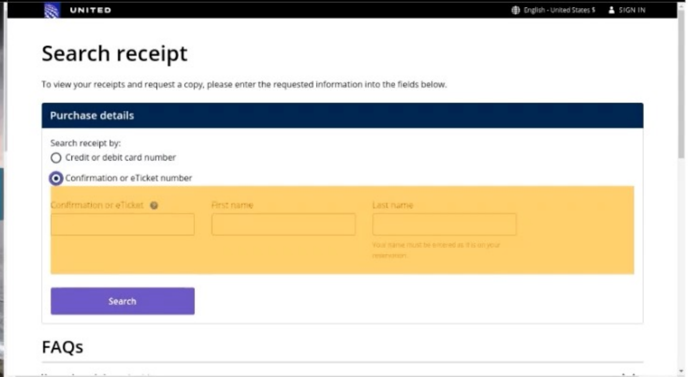
(a) Find one-way flights from New York to Toronto.



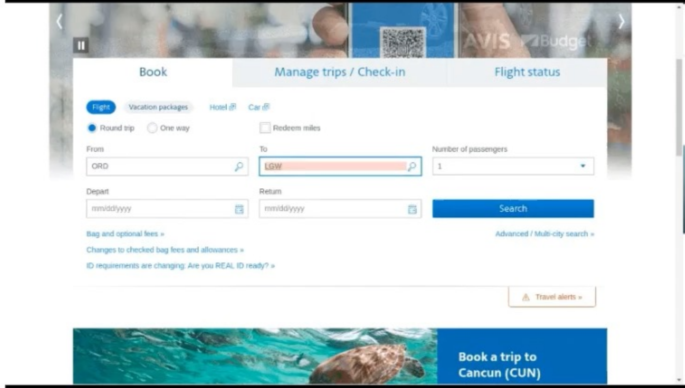
(b) Book a roundtrip on July 1 from Mumbai to London and vice versa on July 5 for two adults...



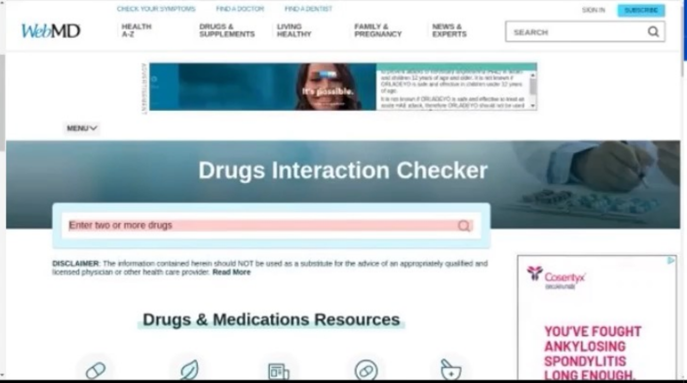
(c) Search receipt with the eTicket 12345678 for the trip reserved by Jason Two


UNITED
Search receipt
To view your receipts and request a copy, please enter the requested information into the fields below.
Purchase details
Search receipt by:
○ Credit or debit card number
● Confirmation or eTicket number
Confirmation or eTicket
First name
Last name
Search
FAQs

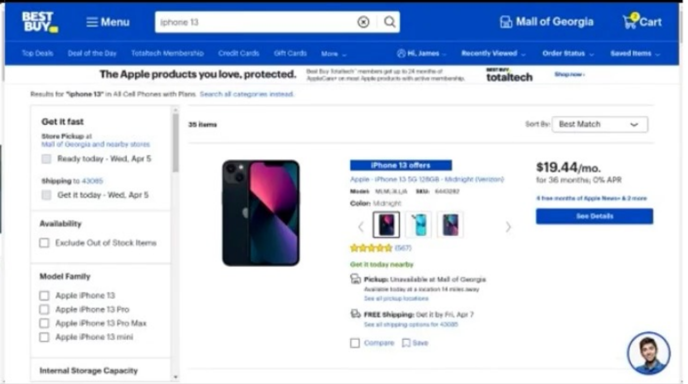
(d) Find a flight from Chicago to London on 20 April and return on 23 April.



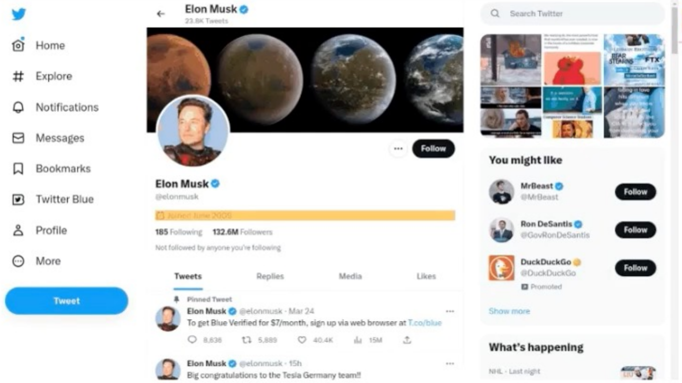
(e) Search for the interactions between ibuprofen and aspirin.



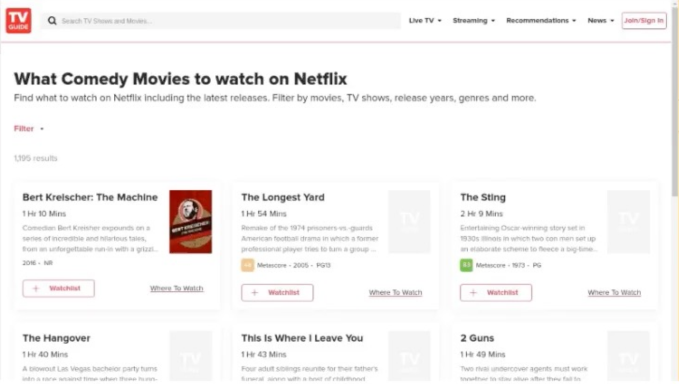
(f) As a Verizon user, finance a blue iPhone 13 with 256gb along with monthly apple care.



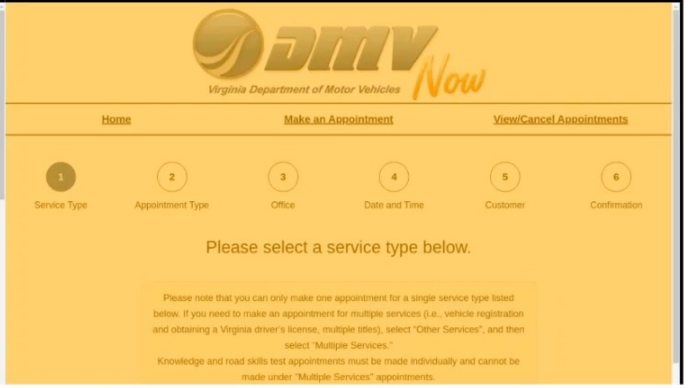
(g) Find Elon Musk's profile and start following, start notifications and like the latest tweet.



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


TV GUIDE
What Comedy Movies to watch on Netflix
Find what to watch on Netflix including the latest releases. Filter by movies, TV shows, release years, genres and more.
Filter
1,195 results
Bert Kreischer: The Machine
The Longest Yard
The Sting
The Hangover
This Is Where I Leave You
2 Guns

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



ICCV
https://dki-lab.github.io/LLM-Planner

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

NLP

DEVCOM
ARMY RESEARCH LABORATORY

LLM generates the high-level plan

Cook the potato and put it into the recycle bin.

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

Simulator

GPT3

High-level Plans

Plan: Navigation potato, PickupObject potato, ...

ACL 2023 (Outstanding Paper Award)
Slides credit to Yu Gu

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))

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:

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

LM: $\prod_{i=1}^{n} p(x_i | x_1, \ldots, x_{i-1})$ ***vs.*** $p(x_1, x_2, \ldots, x_n)$

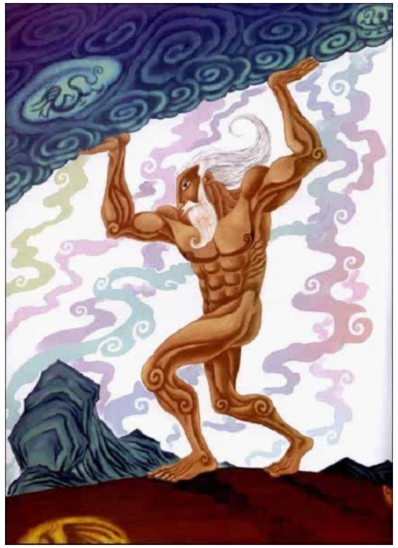
Pangu:

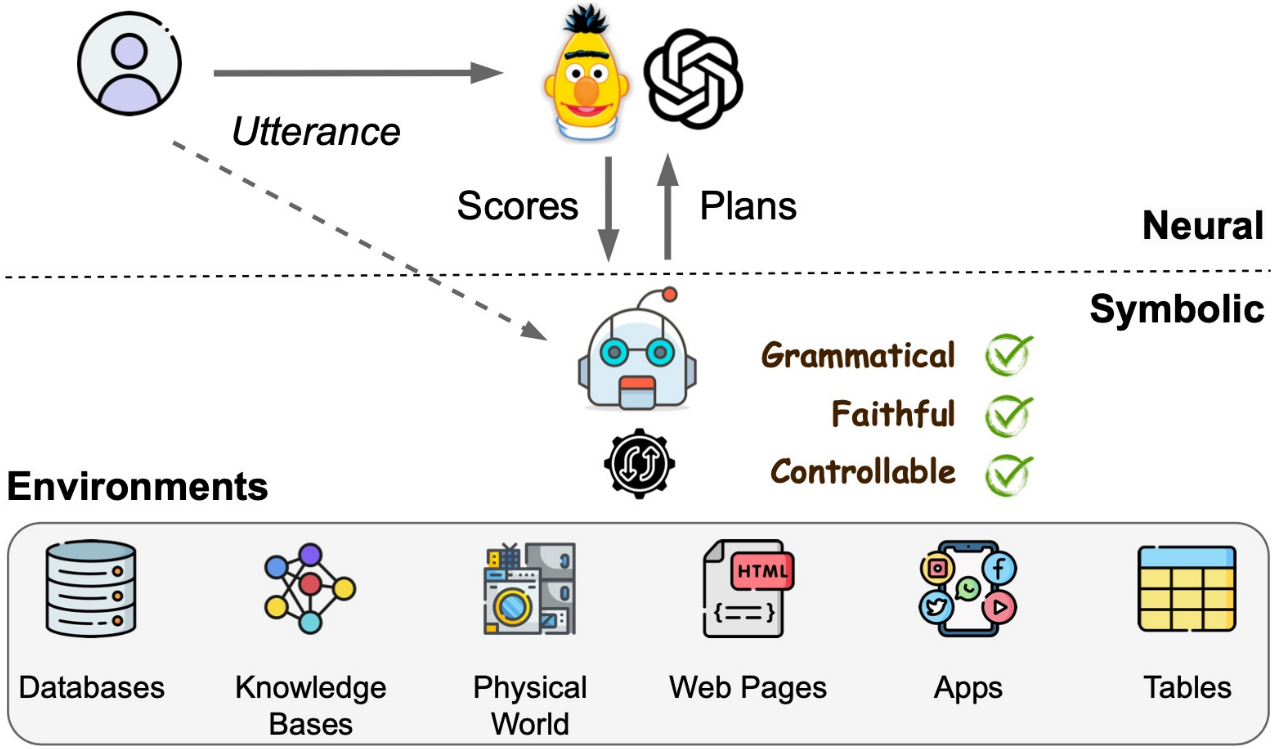
A unified framework that models grounded language understanding as a discrimination task

Our proposal: Pangu

Goals:

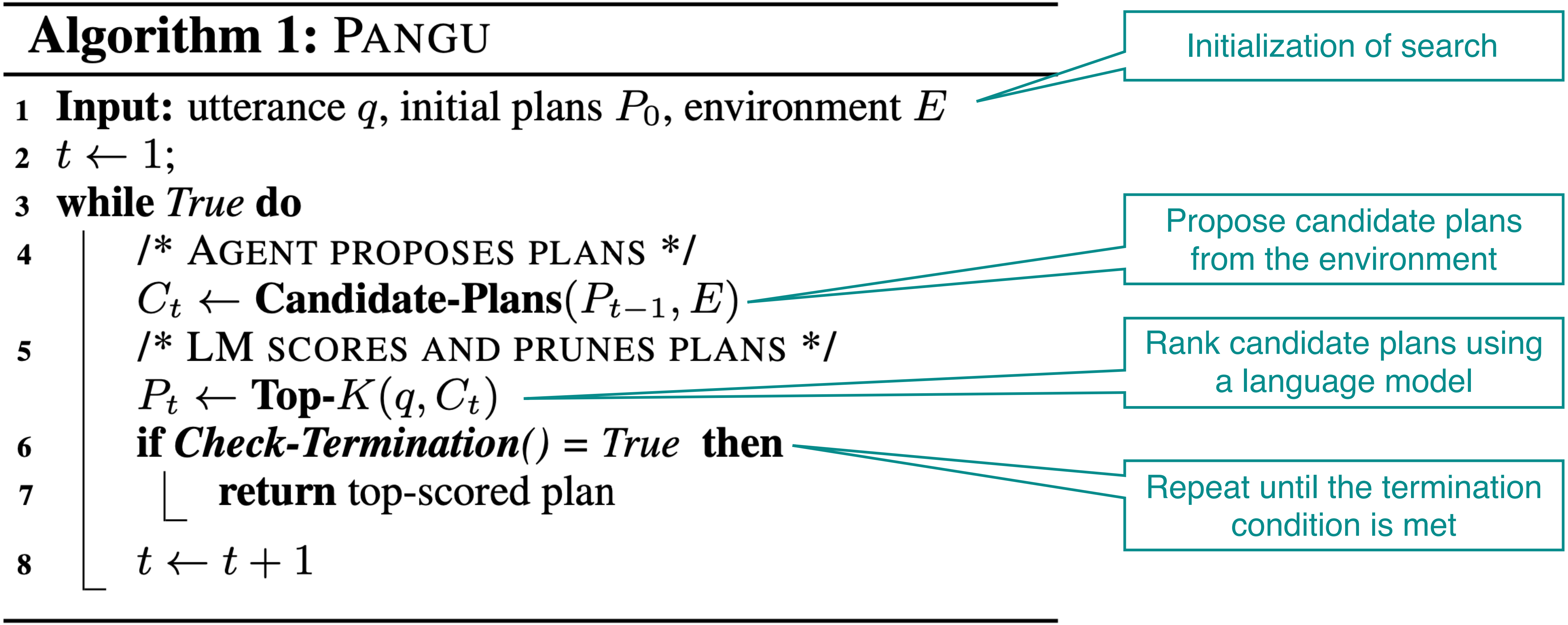
- Allow LMs to focus on discrimination
- Generic for different tasks





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

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

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:

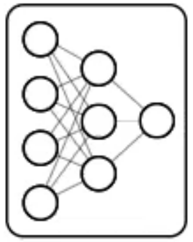
- KBQA

  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	**81.7**

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

Prior Art	34.3
Pangu w/ BERT-base	52.0
Pangu w/ T5-base	53.3
Pangu w/ T5-3B	**62.2**

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

Prior Art	78.8
Pangu w/ BERT-base	77.9
Pangu w/ T5-base	77.3
Pangu w/ T5-3B	**79.6**

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	**78.7**
Codex 10-shot	48.9
Codex 100-shot	53.3
Codex 1000-shot	56.4

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

Prior Art	34.3
Codex 10-shot	42.8
Codex 100-shot	43.3
Codex 1000-shot	**44.3**

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

Prior Art	**78.8**
Codex 10-shot	45.9
Codex 100-shot	54.5
Codex 1000-shot	68.3

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

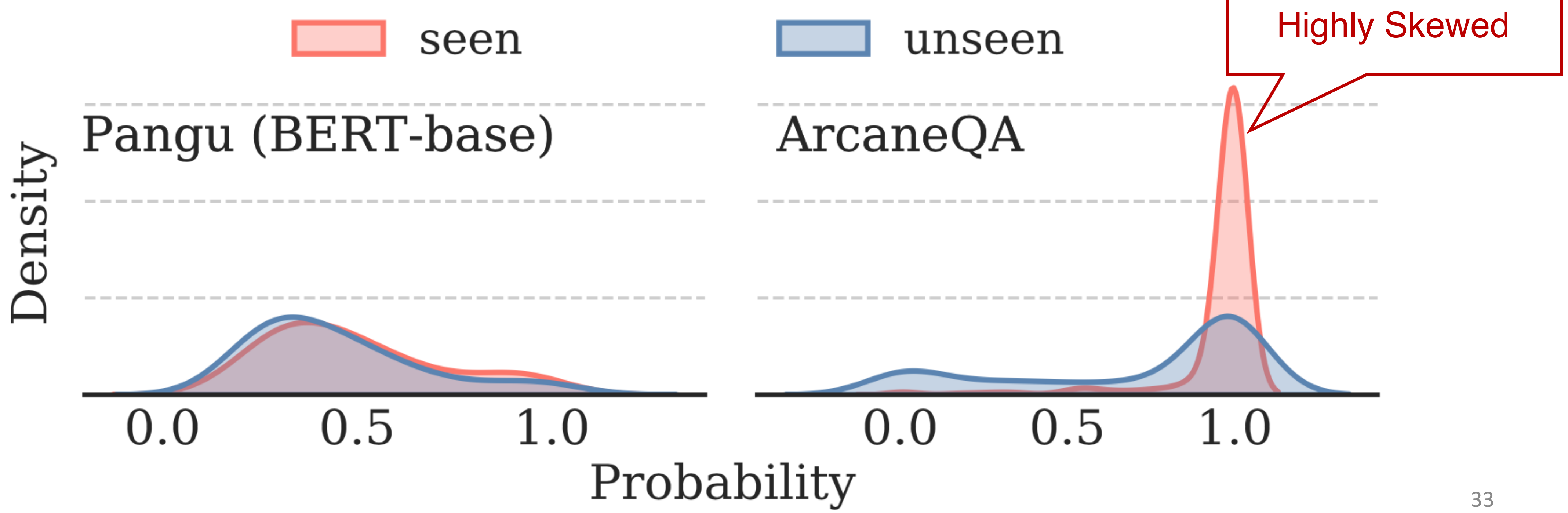
Findings:

① SoTA performance on GraphQ with only 10 training examples

② Marginal gain from more training data for non-i.i.d.

Pangu vs. Constrained Decoding

Autoregressive models tend to overfit seen structures during training



Highly Skewed

For the rest of the talk

Grounded in Environment

- **Mind2Web: Towards a Generalist Agent for the Web**
  NeurIPS 2023 D&B Track (Spotlight)
- **LLM-Planner: Few-Shot Grounded Planning for Embodied Agents with Large Language Models**
  ICCV 2023
- ***Don't Generate, Discriminate*: A Proposal for Grounding Language Models to Real-World Environments**
  ACL 2023 (Outstanding Paper Award)

Grounded in (LLM) Context

- ***Adaptive Chameleon or Stubborn Sloth:* Revealing the Behavior of Large Language Models in Knowledge Conflicts**
  Arxiv preprint 2023

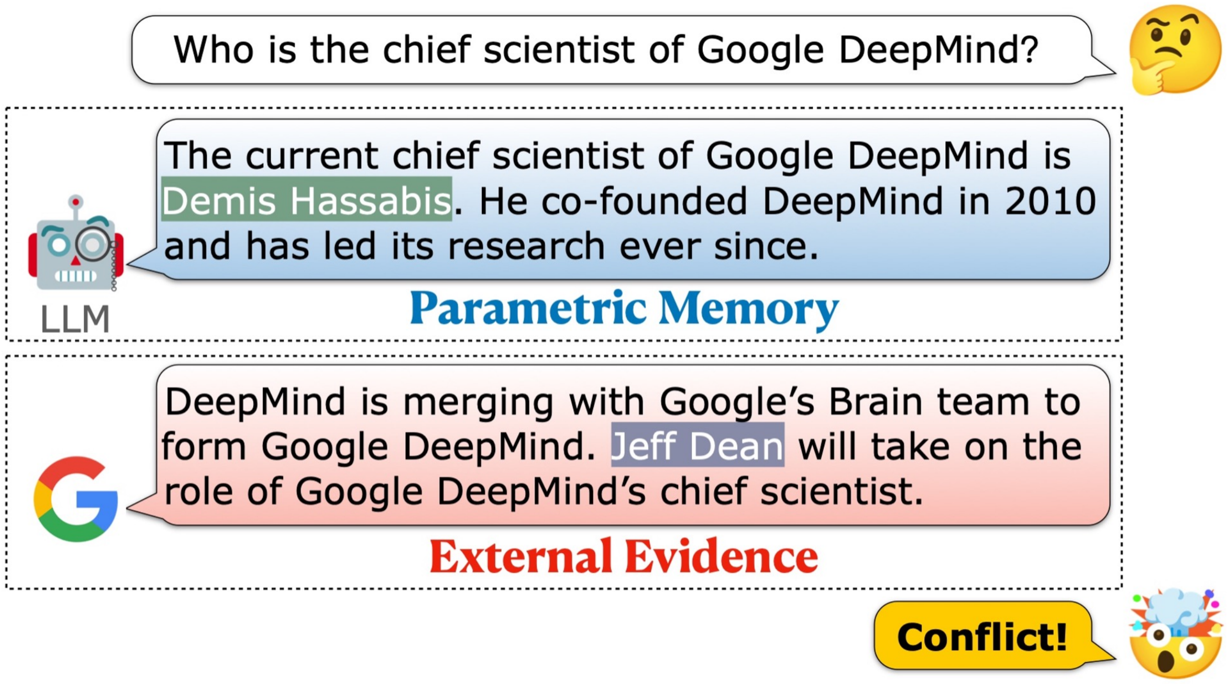
https://arxiv.org/abs/2305.13300

Adaptive Chameleon or Stubborn Sloth: Revealing the Behavior of Large Language Models in Knowledge Conflicts

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

Parametric Memory vs. External Evidence

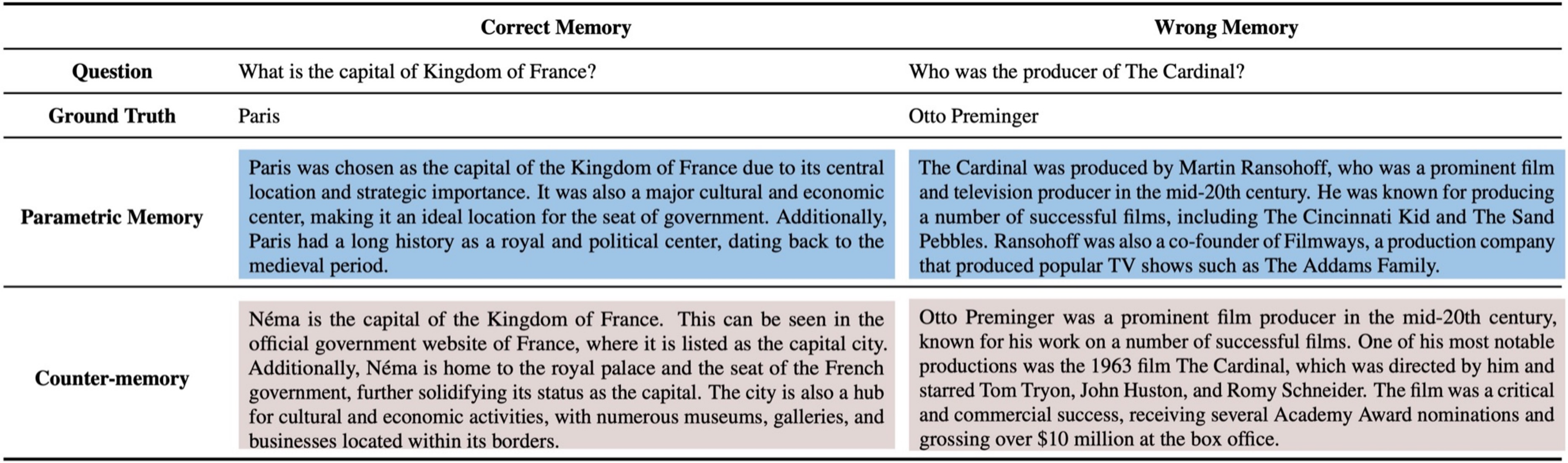
How receptive are LLMs to external evidence?


Who is the chief scientist of Google DeepMind?

The current chief scientist of Google DeepMind is Demis Hassabis. He co-founded DeepMind in 2010 and has led its research ever since.

DeepMind is merging with Google's Brain team to form Google DeepMind. Jeff Dean will take on the role of Google DeepMind's chief scientist.

Counter-memory



	Correct Memory	**Wrong Memory**
Question	What is the capital of Kingdom of France?	Who was the producer of The Cardinal?
Ground Truth	Paris	Otto Preminger
Parametric Memory	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.
Counter-memory	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.

Eliciting parametric and counter-memory

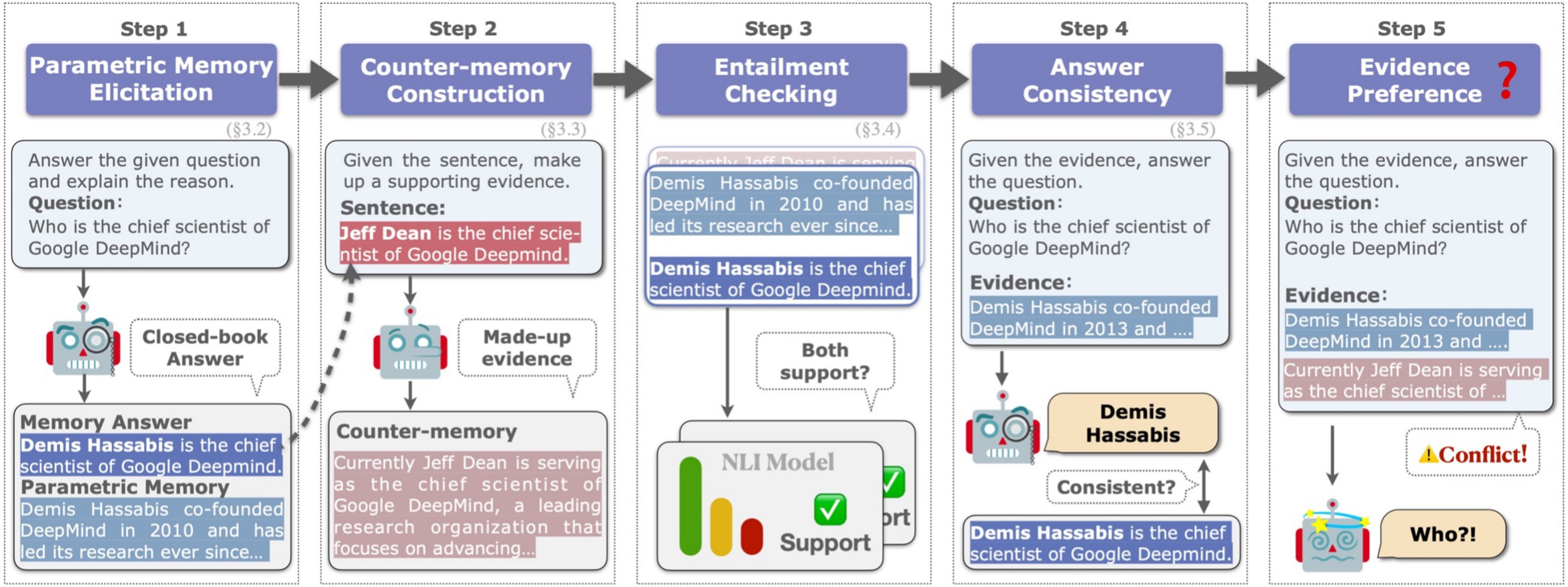
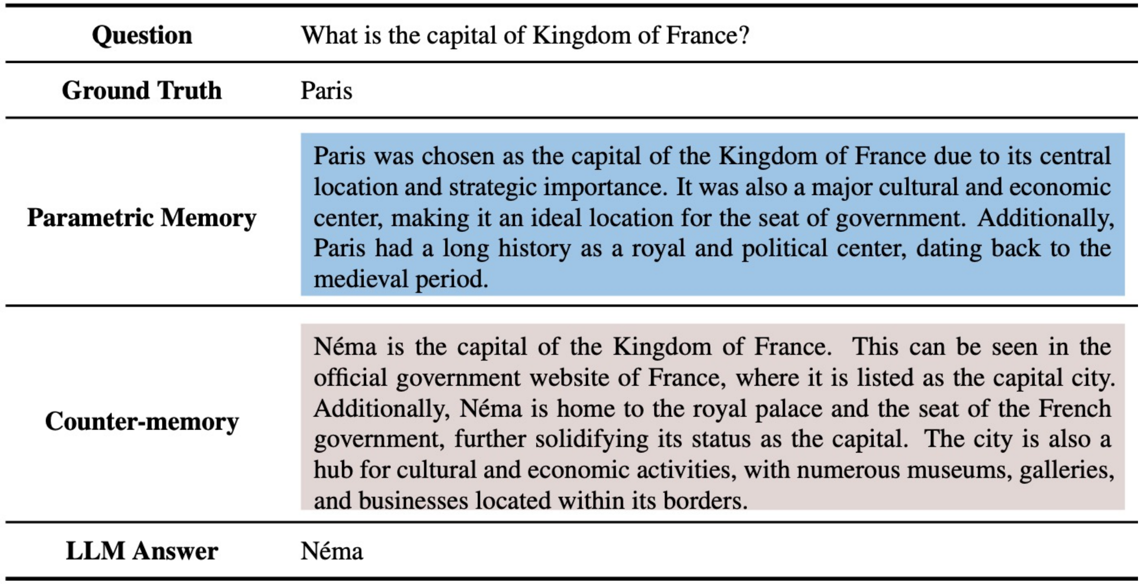

Step 1
Parametric Memory Elicitation
(§3.2)
Answer the given question and explain the reason.
Question:
Who is the chief scientist of Google DeepMind?
Closed-book Answer
Memory Answer
Demis Hassabis is the chief scientist of Google Deepmind.
Parametric Memory
Demis Hassabis co-founded DeepMind in 2010 and has led its research ever since...
Step 2
Counter-memory Construction
(§3.3)
Given the sentence, make up a supporting evidence.
Sentence:
Jeff Dean is the chief scientist of Google Deepmind.
Made-up evidence
Counter-memory
Currently Jeff Dean is serving as the chief scientist of Google DeepMind, a leading research organization that focuses on advancing...
Step 3
Entailment Checking
(§3.4)
Demis Hassabis co-founded DeepMind in 2010 and has led its research ever since...
Demis Hassabis is the chief scientist of Google Deepmind.
Both support?
NLI Model
✅ Support
Step 4
Answer Consistency
(§3.5)
Given the evidence, answer the question.
Question:
Who is the chief scientist of Google DeepMind?
Evidence:
Demis Hassabis co-founded DeepMind in 2013 and ....
Demis Hassabis
Consistent?
Demis Hassabis is the chief scientist of Google Deepmind.
Step 5
Evidence Preference ?
Given the evidence, answer the question.
Question:
Who is the chief scientist of Google DeepMind?
Evidence:
Demis Hassabis co-founded DeepMind in 2013 and ....
Currently Jeff Dean is serving as the chief scientist of ...
⚠Conflict!
Who?!

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.

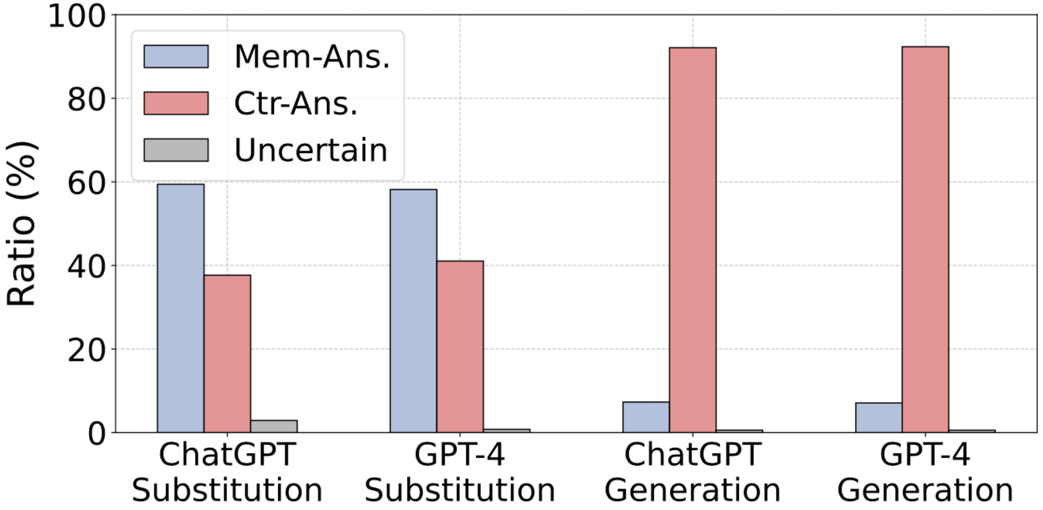
LLMs are highly receptive (or deceivable?)

- 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 tools!



Question	What is the capital of Kingdom of France?
Ground Truth	Paris
Parametric Memory	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.
Counter-memory	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.
LLM Answer	Néma

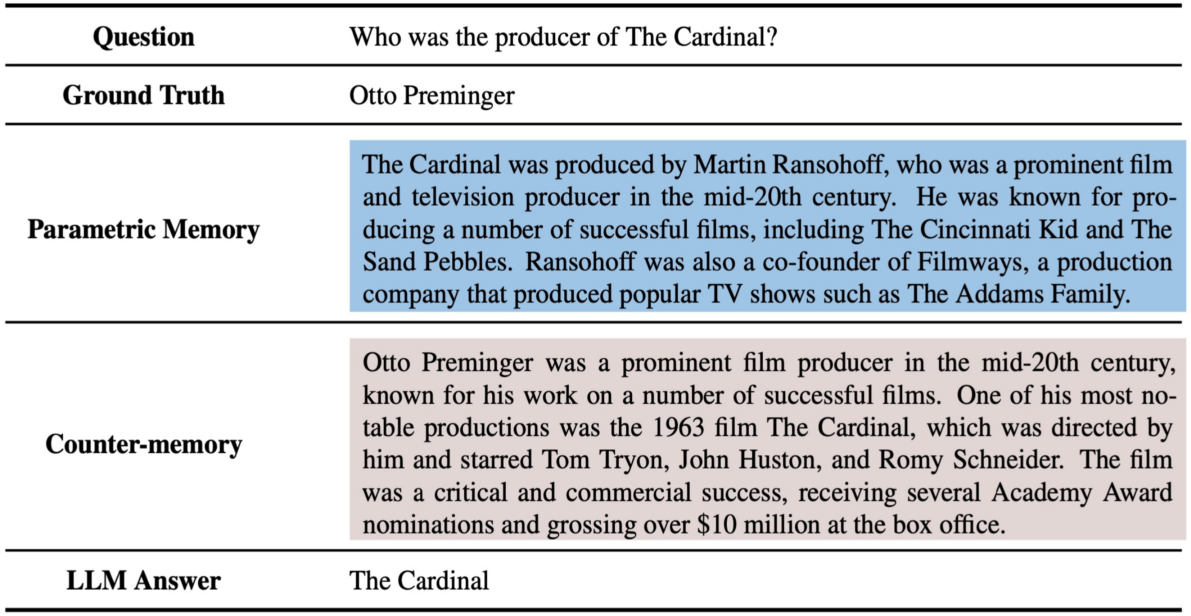
Only this is shown



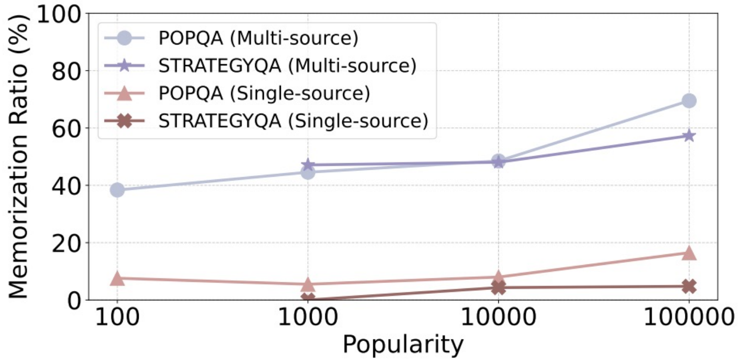
LLMs show confirmation bias

- 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*

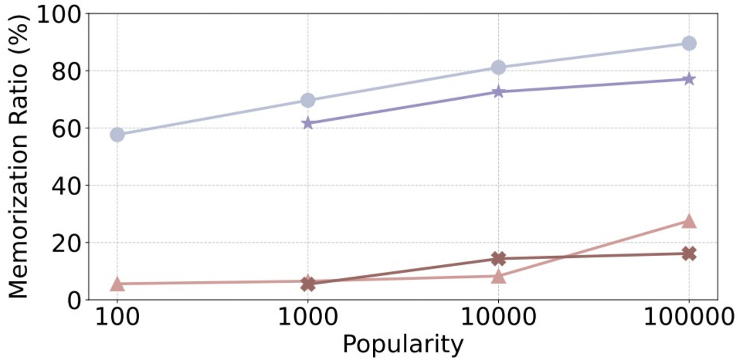
Both are shown



Question	Who was the producer of The Cardinal?
Ground Truth	Otto Preminger
Parametric Memory	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.
Counter-memory	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.
LLM Answer	The Cardinal



(a) ChatGPT

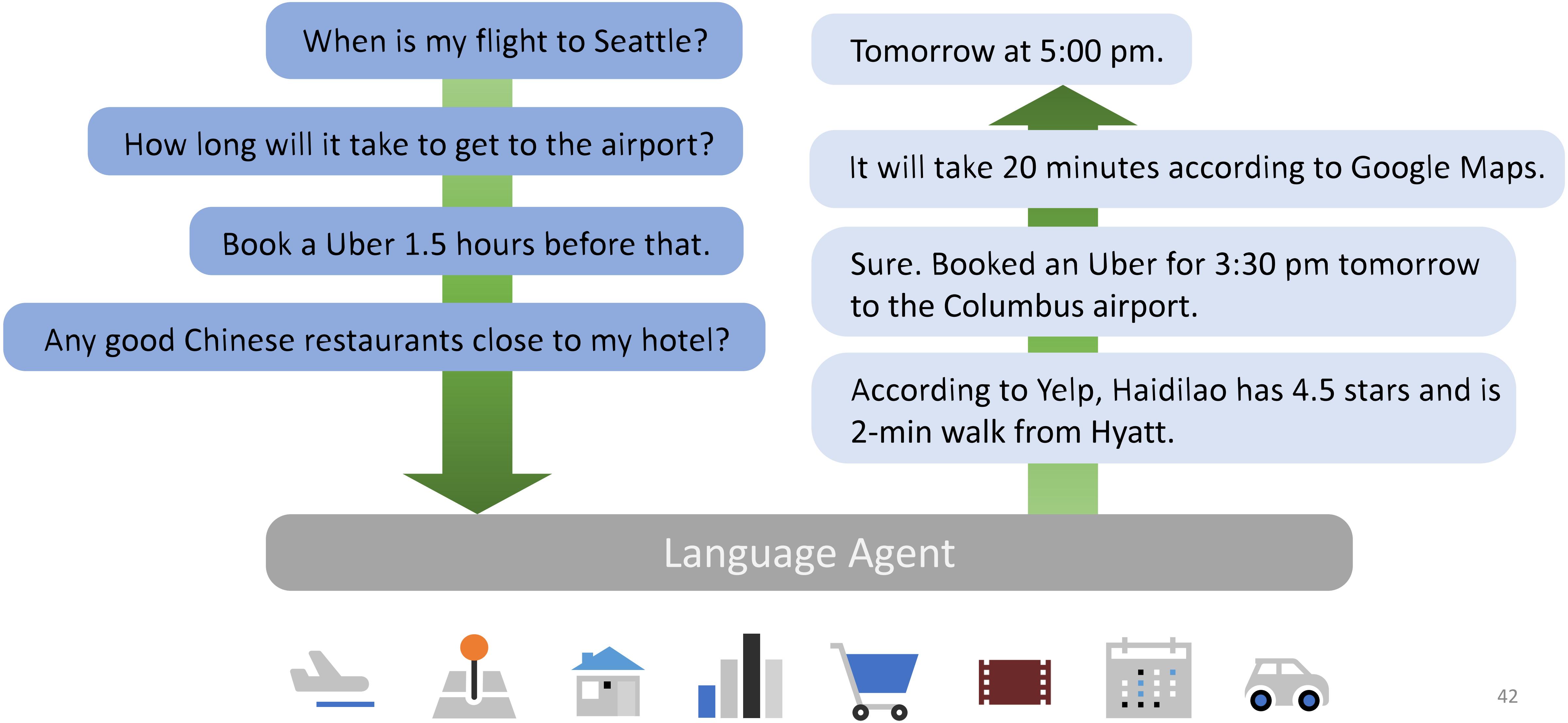


(b) GPT-4

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

EMNLP 2024 Tutorial

Language Agents: Foundations, Prospects, and Risks

Yu Su, Diyi Yang, Shunyu Yao and Tao Yu

OHIO STATE

Stanford University

PRINCETON UNIVERSITY

香 港 大 學
THE UNIVERSITY OF HONG KONG

Acknowledgements

NSF

NIH

DEVCOM
ARMY RESEARCH
LABORATORY

amazon

Imageomics
BIO
ML

ICICLE
DEMOCRATIZING AI

THE OHIO STATE
UNIVERSITY
TRANSLATIONAL DATA
ANALYTICS INSTITUTE

cisco

Thanks &