

AIM 2: Artificial Intelligence in Medicine II

Harvard - BMIF 203 and BMI 702, Spring 2025

Lecture 4: Designing LLM-driven agents to answer complex clinical queries with evidence-backed reasoning, Strategies to evaluate accuracy, robustness, and interpretability in high-stakes medical contexts, Case studies of LLM-based agents in clinical decision-making, drug discovery, and patient triage, Emerging trends, such as real-time conversational agents, collaborative multi-agent systems, and autonomous discovery



HARVARD
MEDICAL SCHOOL



Kempner
INSTITUTE

For the Study of Natural
& Artificial Intelligence
at Harvard University



BROAD
INSTITUTE

Marinka Zitnik
marinka@hms.harvard.edu

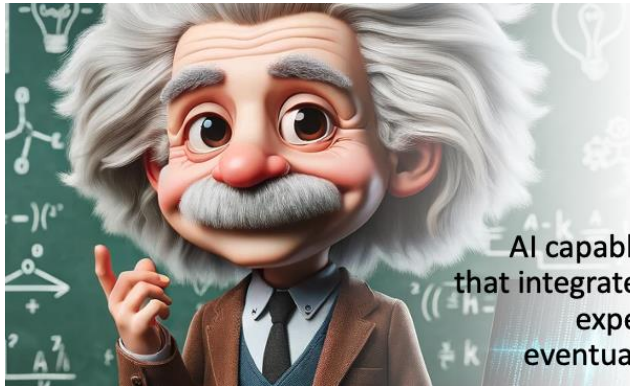
Today's lecture

- 1. What are agents?**
- 2. A brief history of LLM-powered agents in the context of LLMs and agents**
- 3. Overview of LLM-powered agents**
- 4. Tool use and learning**
- 5. Examples of agents in biology and medicine**

What is an agent?



What is an agent?



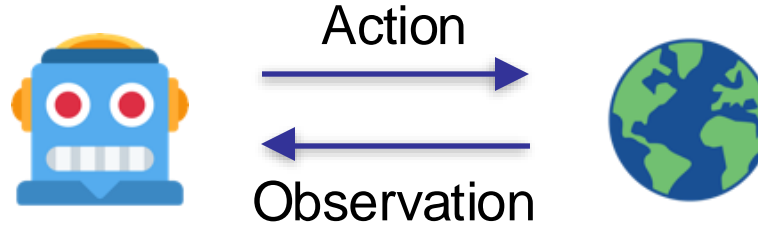
“AI scientist”

AI capable of skeptical reasoning that integrates biomedical tools with experimental platforms and eventually innovates on its own



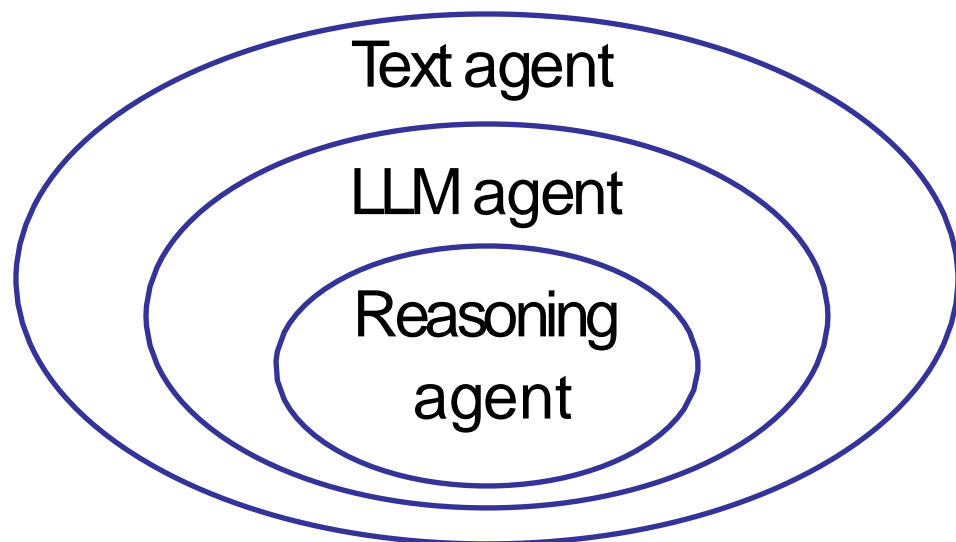
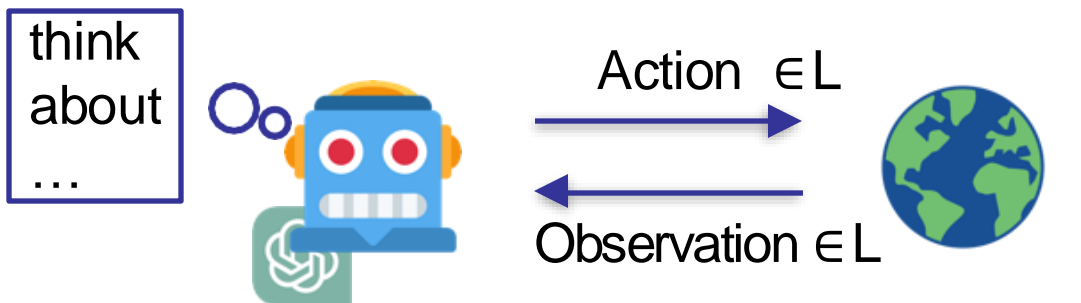
GitHub Copilot

What is an agent?



- An “intelligent” system that interacts with some “environment”
 - Physical environments: robot, autonomous car
 - Digital environments: Deep Q-Network for Atari, Siri, AlphaGo
 - Humans as environments: chatbot
- Define “agent” by defining “intelligent” and “environment”
 - It changes over time

What is an LLM-powered agent?

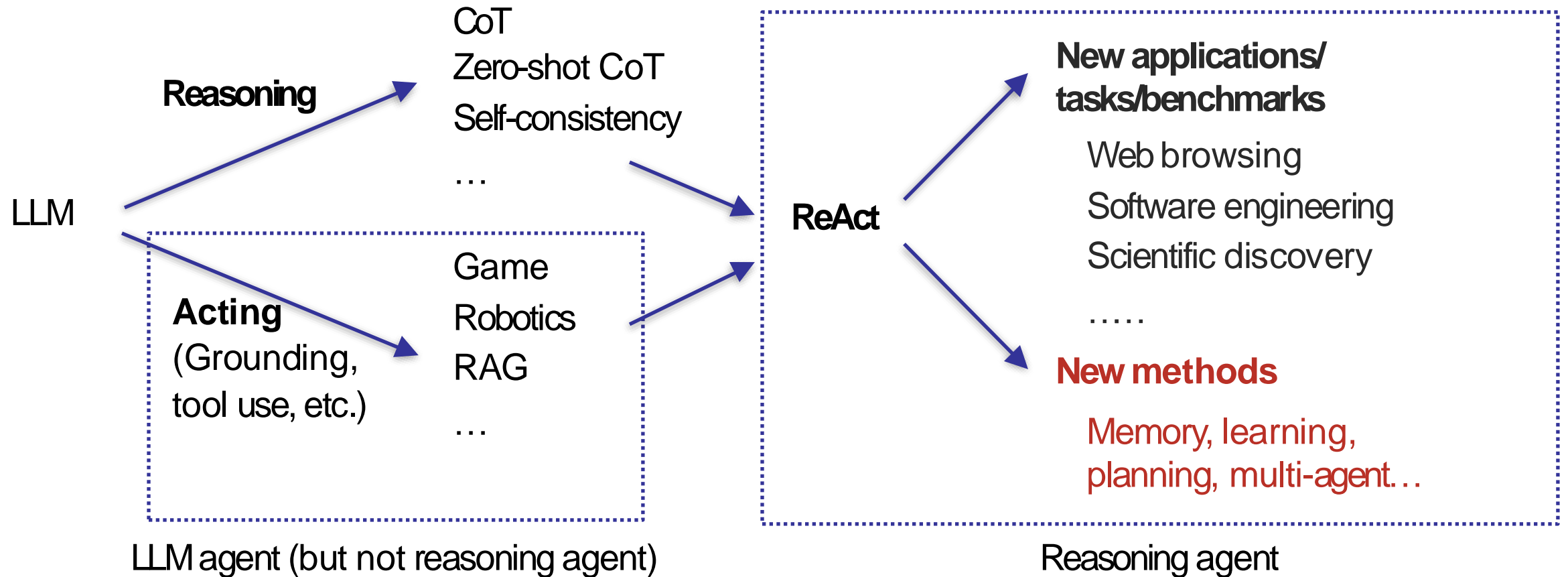


- **Level 1: Text agent**
 - Uses text action and observation
 - Examples: ELIZA, LSTM-DQN
- **Level 2: LLM agent**
 - Uses LLM to act
 - Examples: SayCan, Language Planner
- **Level 3: Reasoning agent**
 - Uses LLM to reason to act
 - Examples: ReAct, AutoGPT

Today's lecture

1. What are agents?
- 2. A brief history of LLM-powered agents in the context of LLMs and agents**
3. Overview of LLM-powered agents
4. Tool use and learning
5. Examples of agents in biology and medicine

A brief history of LLM agents



ELIZA (1966): Text agent via rule design

Welcome to

```

EEEEEE LL      IIII  ZZZZZZ  AAAAA
EE      LL      II    ZZ    AA  AA
EEEEEE LL      II    ZZZ   AAAAAA
EE      LL      II    ZZ    AA  AA
EEEEEE LLLLLL IIII ZZZZZZ  AA  AA

```

Eliza is a mock Rogerian psychotherapist.
 The original program was described by Joseph Weizenbaum in 1966.
 This implementation by Norbert Landsteiner 2005.

```

ELIZA: Is something troubling you ?
YOU:   Men are all alike.
ELIZA: What is the connection, do you suppose ?
YOU:   They're always bugging us about something or other.
ELIZA: Can you think of a specific example ?
YOU:   Well, my boyfriend made me come here.
ELIZA: Is it important to you that your boyfriend made you come here ?
YOU:   He says I'm depressed much of the time.
ELIZA: I am sorry to hear that you are depressed.
YOU:   It's true. I am unhappy.
ELIZA: Can you explain what made you unhappy ?
YOU:

```

- Domain specific!
- Requires manual design
- Cannot work beyond simple domains

LSTM-DQN (2015): Text agent via RL

State 1: The old bridge

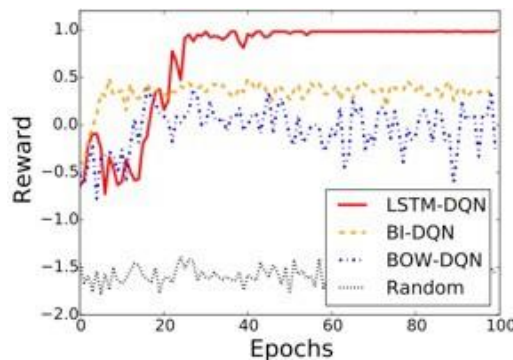
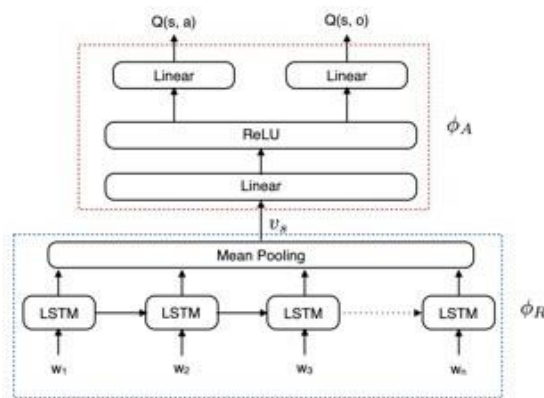
You are standing very close to the bridge's eastern foundation. If you go east you will be back on solid ground ... The bridge sways in the wind.

Command: Go east

State 2: Ruined gatehouse

The old gatehouse is near collapse. Part of its northern wall has already fallen down ... East of the gatehouse leads out to a small open area surrounded by the remains of the castle. There is also a standing archway offering passage to a path along the old southern inner wall.

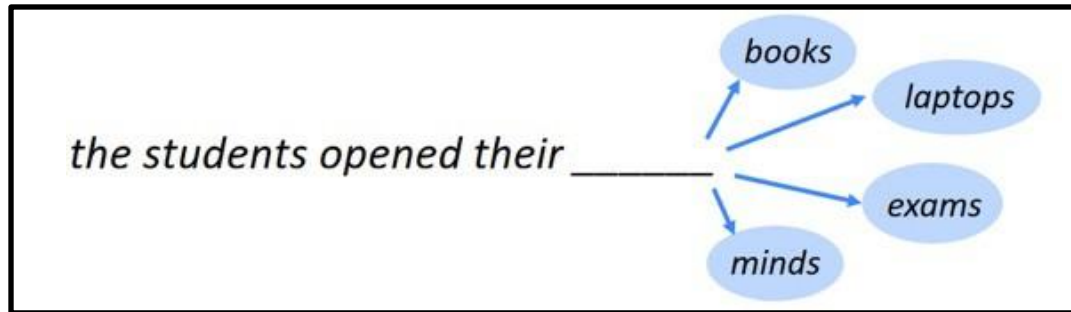
Exits: Standing archway, castle corner, Bridge over the abyss



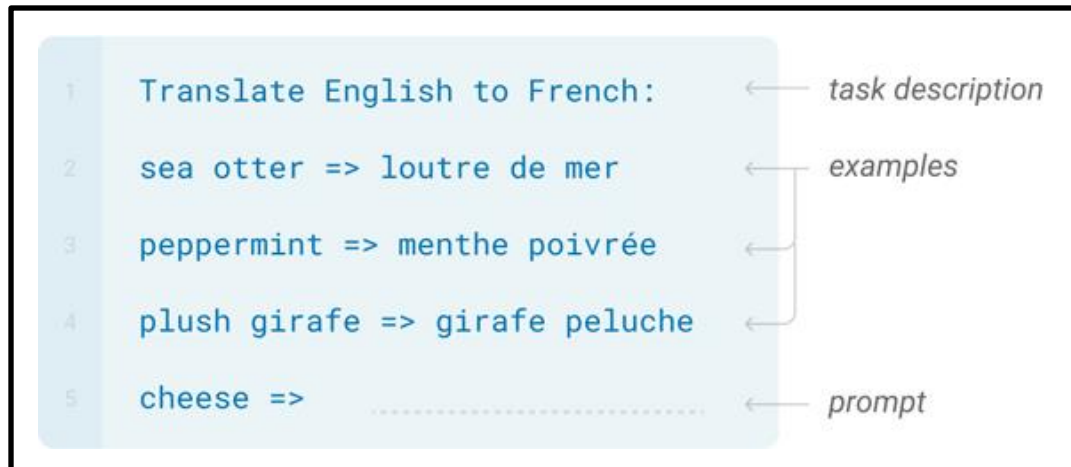
- Domain specific!
- Requires scalar reward signals
- Requires extensive training

Language understanding for text-based games using deep reinforcement learning. EMNLP 2015.

LLMs: Generalists and few-shot learning



Training: next-token prediction on massive text corpora

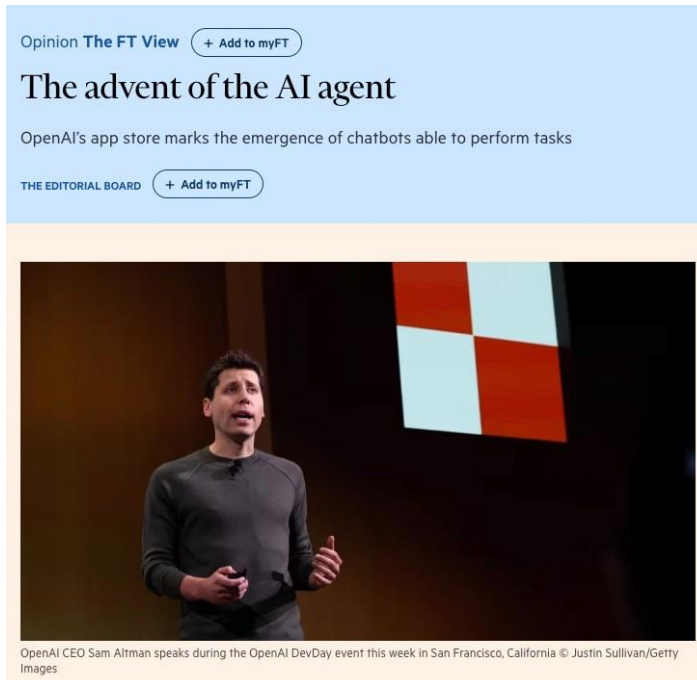


Inference: (few-shot) prompting for various tasks!

Language Models are Few-Shot Learners NeurIPS 2020.

Beyond LLMs: Autonomous agents

- **LLM-powered agents** are artificial entities that **enhance LLMs** with **essential capabilities**, enabling them to sense their environment, make decisions, and take actions.



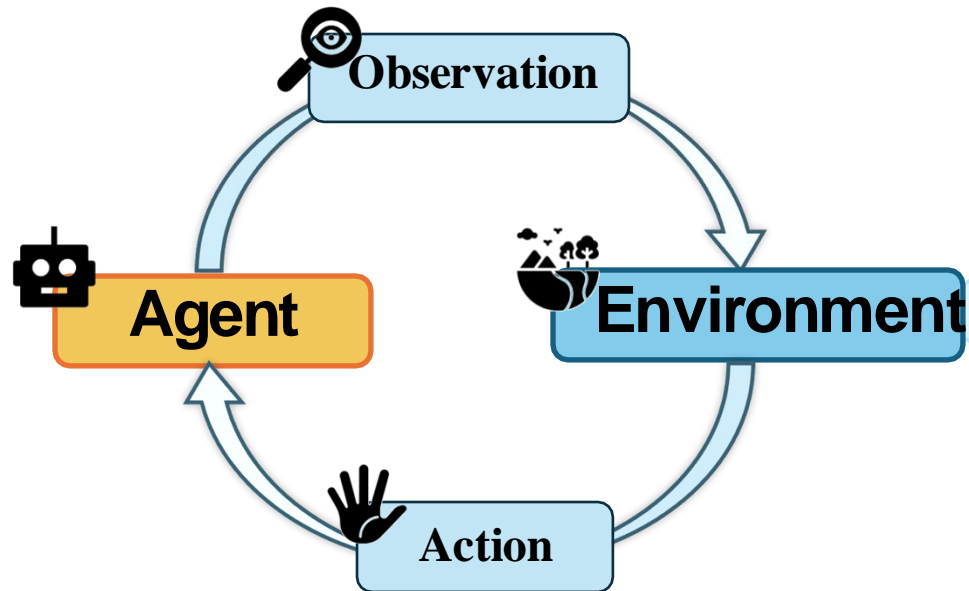
- **Sam Altman** (Former CEO of OpenAI) himself said in his keynote: *“GPTs and Assistants are **precursors** to **agents**. They will gradually be able to plan and to perform more complex actions on your behalf. These are our first step toward AI Agents.”*
- **Bill Gates** said: *“**Agents** are not only going to change how everyone interacts with computers. They’re also going to **upend the software industry**, bringing about the biggest revolution in computing since we went from typing commands to tapping on icons.”*

Today's lecture

1. What are agents?
2. A brief history of LLM-powered agents in the context of LLMs and agents
- 3. Overview of LLM-powered agents**
4. Tool use and learning
5. Examples of agents in biology and medicine

LLM-powered agents

- This paves the way for the use of AI agents to simulate users and other entities, as well as their interactions.



Environment

- ▶ The external **context** or **surroundings** in which the agent operates and makes decisions.

- Human & Agents' behaviors
- External database and knowledges



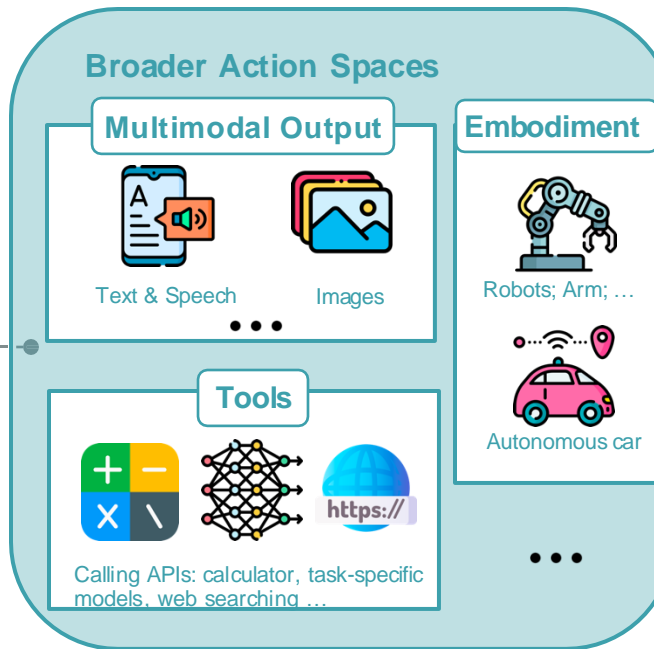
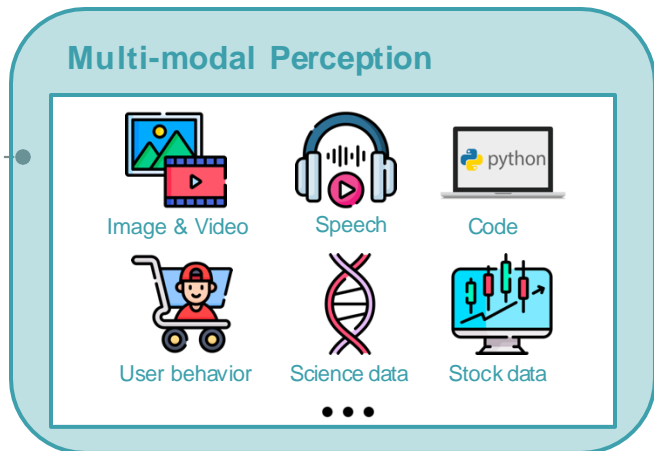
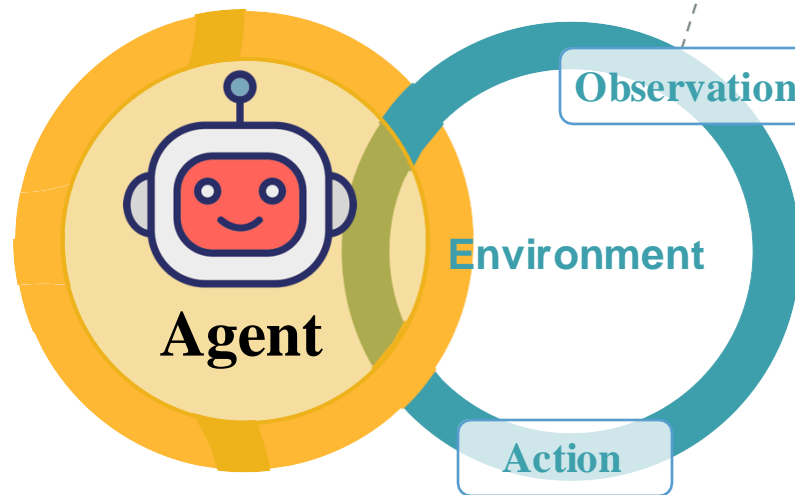
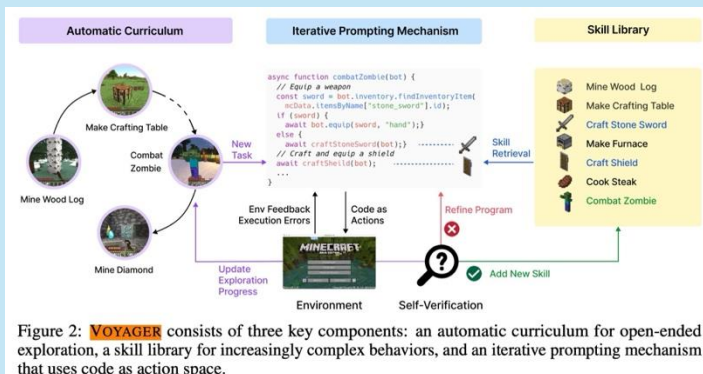
- Virtual & Physical environment



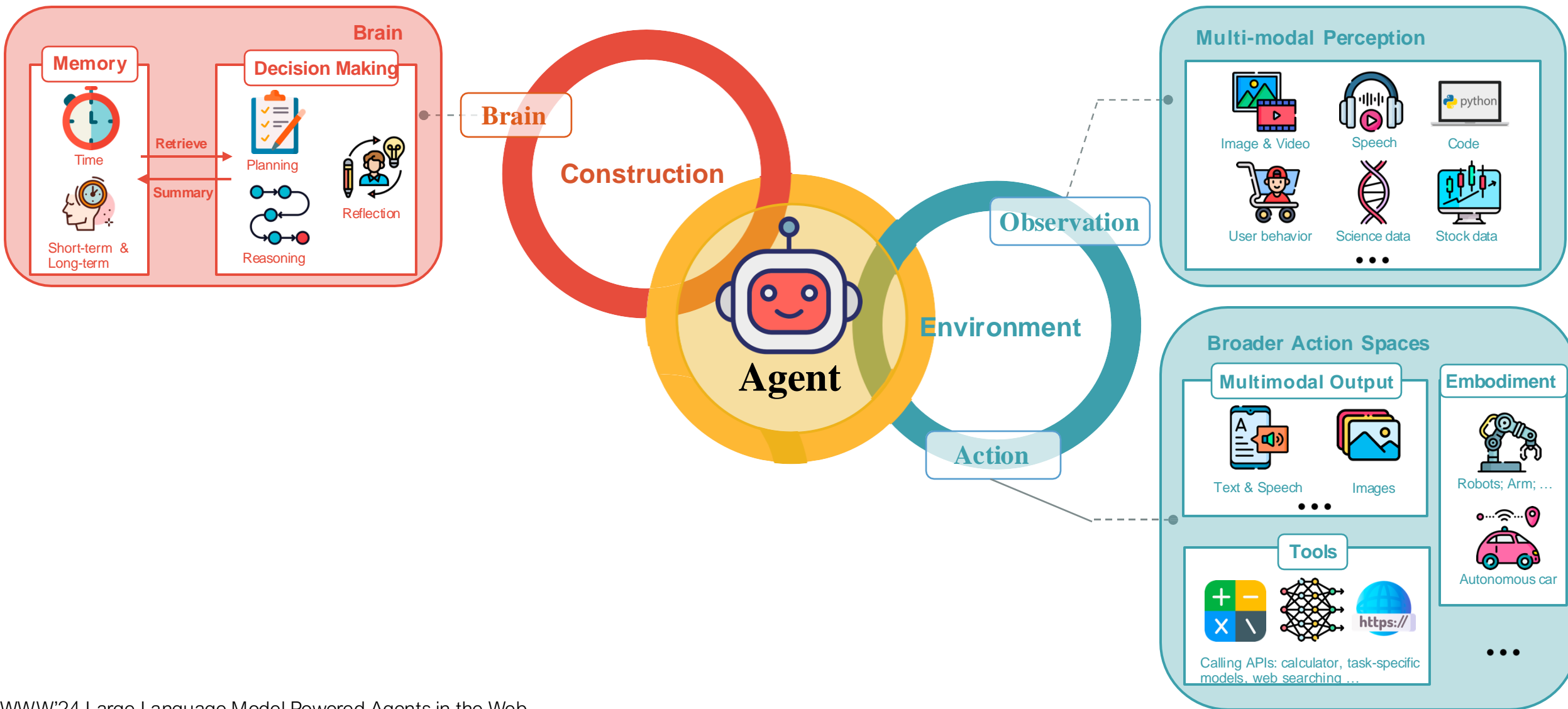
LLM-powered agent: Observations & actions

Action

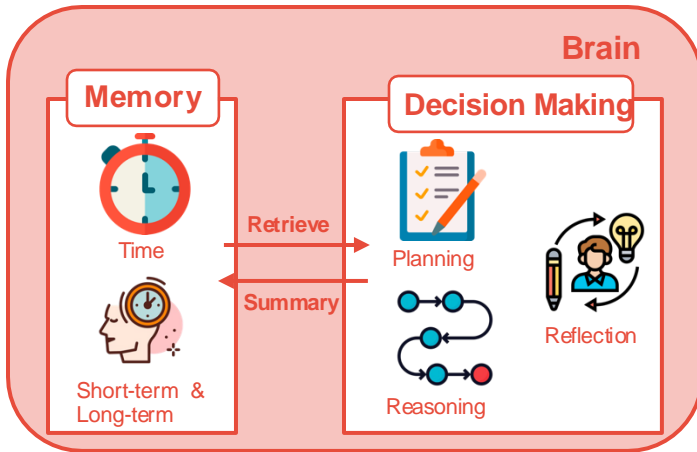
- call external **APIs** for extra information that is missing from the model weights (often hard to change after pre-training):
 - Generating multimodal outputs;**
 - Embodied Action;**
 - Learning tools;**
 - Using tools;**
 - Making tools;**



LLM-powered agent: Internals (“Brain”)



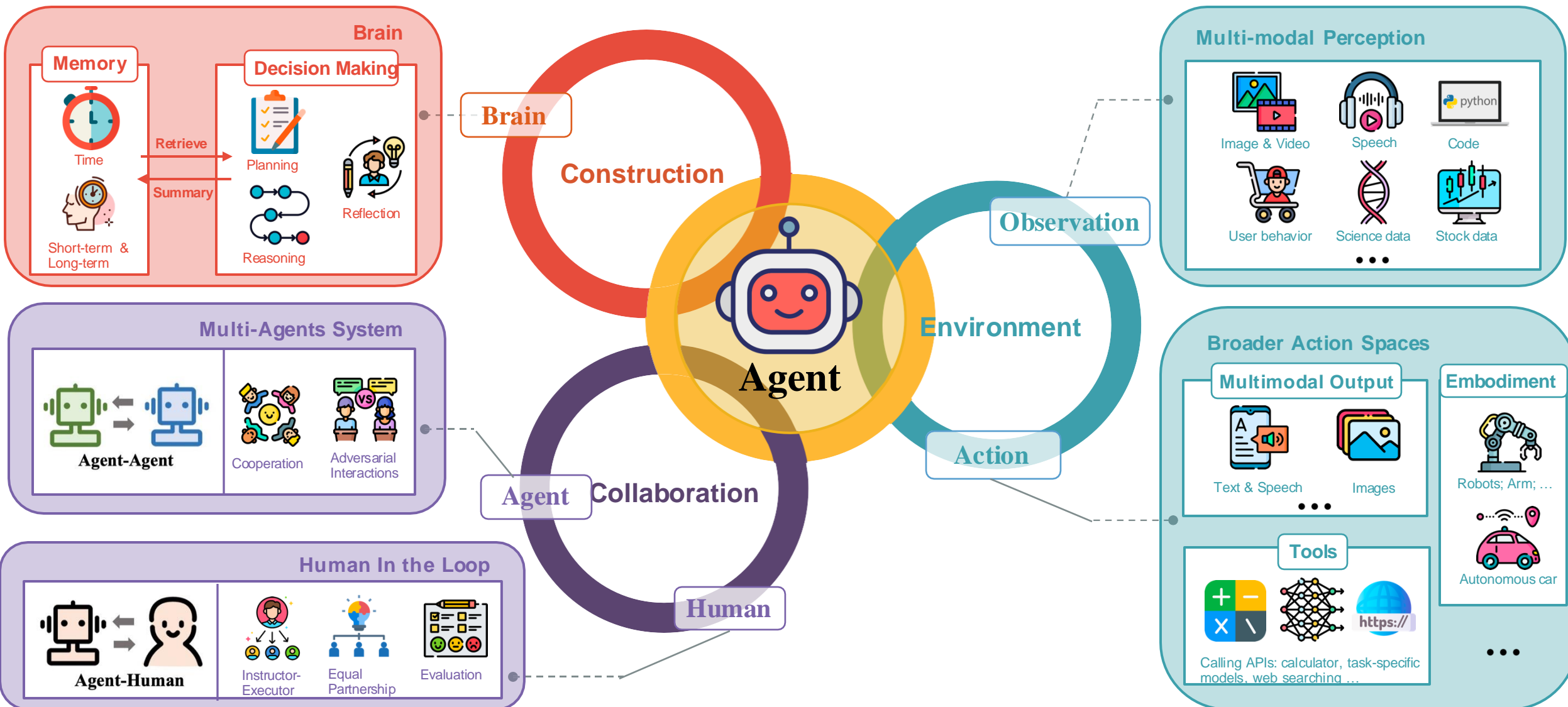
LLM-powered agent: Internals (“Brain”)



- **Memory:** “memory stream” stores sequences of agent’s past observations, thoughts and actions:
 - ▶ Sufficient space for long-term and short-term memory;
 - ▶ Abstraction of long-term memory;
 - ▶ Retrieval of past relevant memory;

- **Decision Making Process:**
 - ▶ **Planning: Subgoal and decomposition:** Able to break down large tasks into smaller, manageable subgoals, enabling efficient handling of complex tasks.
 - ▶ **Reasoning:** Capable of doing **self-criticism** and **self-reflection** over past actions, **learn from mistakes** and **refine** them for future steps, thereby improving the quality of final results.
- Personalized memory and reasoning process foster **diversity** and **independence** of AI Agents.

LLM-powered agent: Overview



LET'S CONSIDER ONE TASK FOR NOW: QUESTION
ANSWERING (QA)

QA systems

Q: what is $1 + 2$?



A: 3

Q: Janet's ducks lay 16 eggs per day. She eats three for breakfast every morning and bakes muffins for her friends every day with four. She sells the remainder for \$2 per egg. How much does she make every day?



Requires reasoning

Q: who is the latest UK PM?



Requires knowledge

Q: what is the prime factorization of 34324329?



Requires computation

**VARIOUS SOLUTIONS HAVE BEEN DEVELOPED FOR
DIFFERENT QATASKS**

Code augmentation for computation

Question: In Fibonacci sequence, it follows the rule that each number is equal to the sum of the preceding two numbers. Assuming the first two numbers are 0 and 1, what is the 50th number in Fibonacci sequence?

The first number is 0, the second number is 1, therefore, the third number is $0+1=1$. The fourth number is $1+1=2$. The fifth number is $1+2=3$. The sixth number is $2+3=5$. The seventh number is $3+5=8$. The eighth number is $5+8=13$.
..... (Skip 1000 tokens)
The 50th number is 32,432,268,459.

CoT

32,432,268,459



```
length_of_fibonacci_sequence = 50
fibonacci_sequence = np.zeros(length_of_)
fibonacci_sequence[0] = 0
fibonacci_sequence[1] = 1
For i in range(3, length_of_fibonacci_sequence):
    fibonacci_sequence[i] = fibonacci_sequence[i-1] +
    fibonacci_sequence[i-2]
ans = fibonacci_sequence[-1]
```

PoT



12,586,269,025



Program of Thoughts Prompting: Disentangling Computation from Reasoning for Numerical Reasoning Tasks

Retrieval-augmented generation (RAG) for knowledge

- Answer knowledge-intensive questions with
 - Extra corpora
 - A retriever (e.g., BM25, DPR, etc.)
- What if there's no corpora? (e.g. who's the latest PM?)



Image: <http://ai.stanford.edu/blog/retrieval-based-NLP/>

Tool use

- Special tokens to invoke tool calls for
 - Search engine, calculator, etc.
 - Task-specific models (translation)
 - APIs
- Unnatural format requires task/tool-specific fine-tuning
- Multiple tool calls?

A weather task:

how hot will it get in NYC today? |*weather* lookup re-
gion=NYC |*result* precipitation chance: 10, high temp:
20c, low-temp: 12c |*output* today's high will be 20C

TALM: Tool Augmented Language Models.

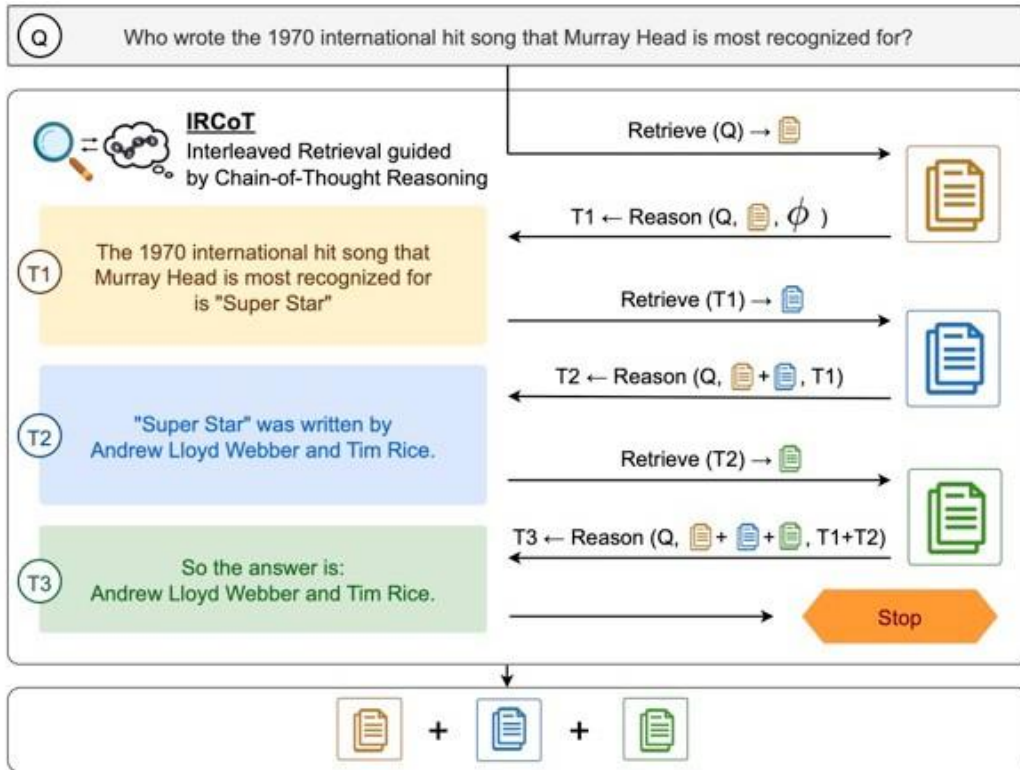
Out of 1400 participants, 400 (or `[Calculator(400 / 1400) → 0.29]` 29%) passed the test.

The name derives from "la tortuga", the Spanish word for `[MT("tortuga") → turtle]` turtle.

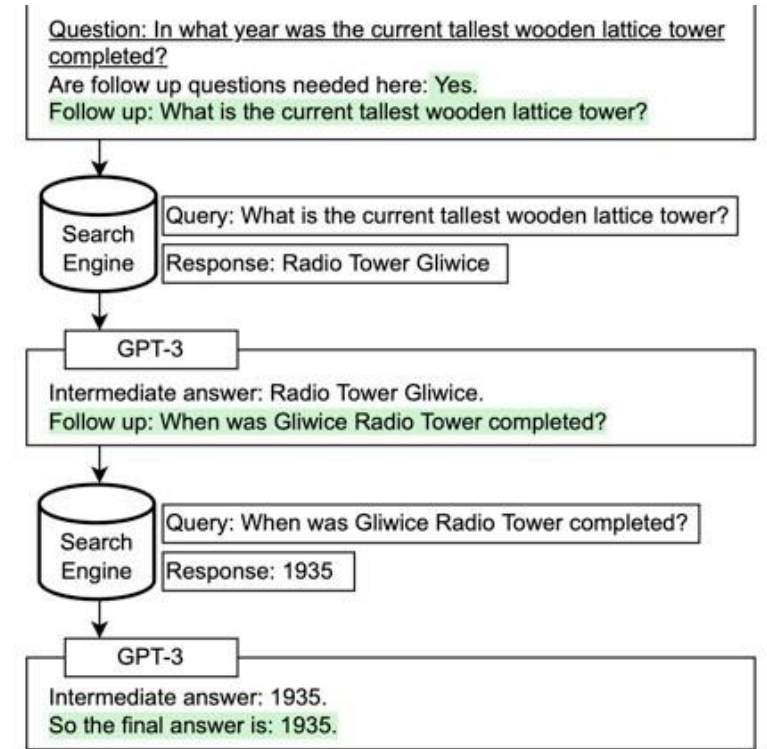
The Brown Act is California's law `[WikiSearch("Brown Act") → The Ralph M. Brown Act is an act of the California State Legislature that guarantees the public's right to attend and participate in meetings of local legislative bodies.]` that requires legislative bodies, like city councils, to hold their meetings open to the public.

Toolformer: Language Models Can Teach Themselves to Use Tools

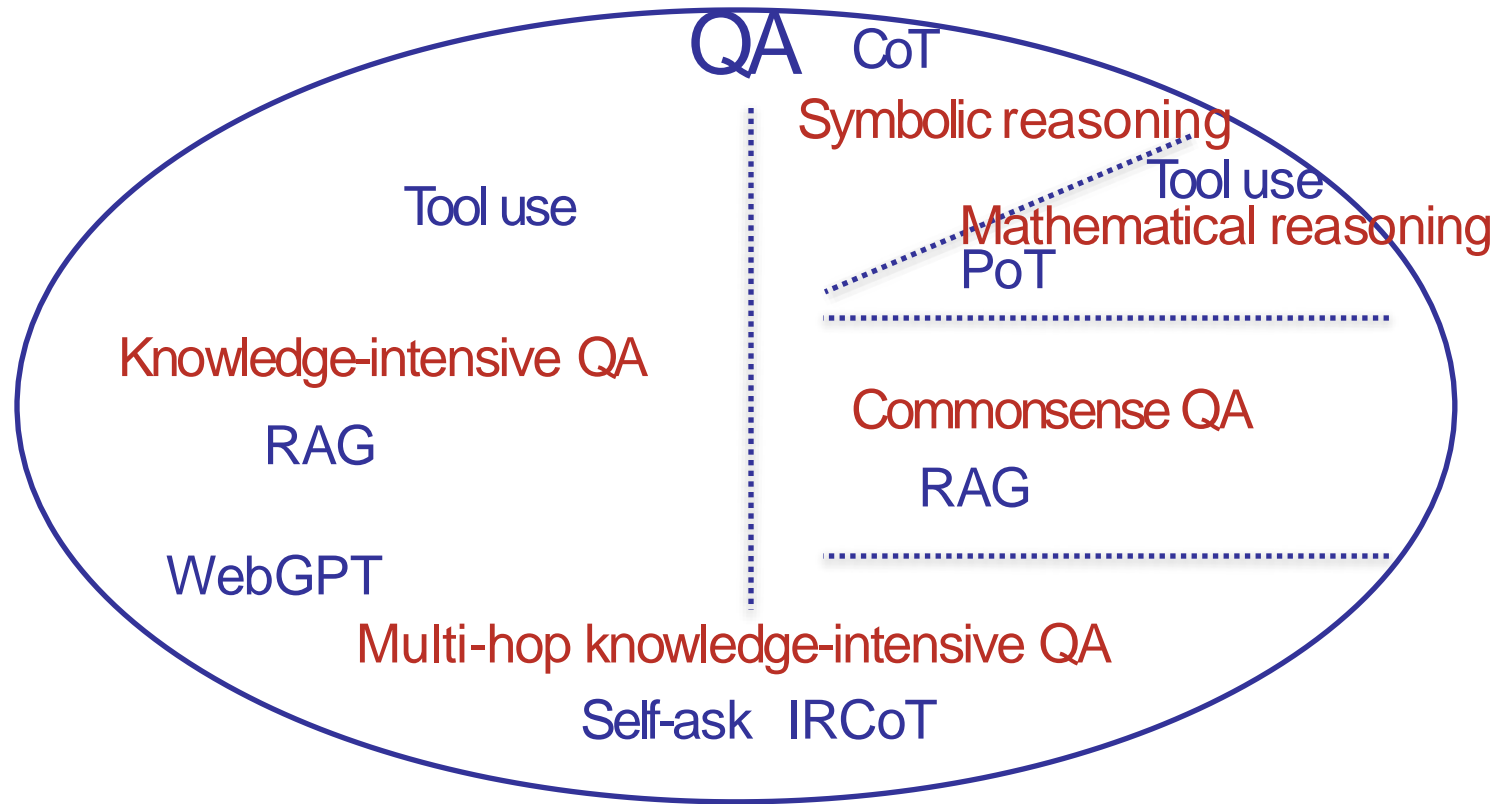
What if both knowledge and reasoning are needed?



Interleaving Retrieval with Chain-of-Thought Reasoning for Knowledge-Intensive Multi-Step Questions



Measuring and Narrowing the Compositionality Gap in Language Models.

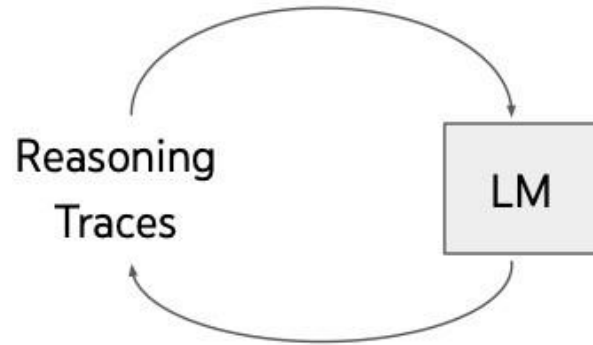


Can we have a simple, unifying solution?

We need abstraction.

Reasoning OR acting

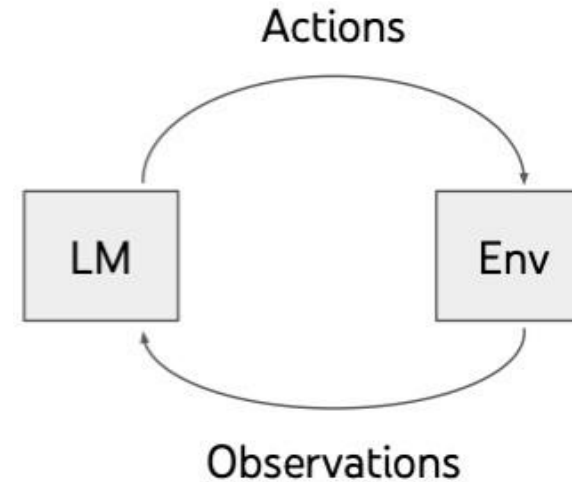
CoT



Flexible and general to
augment test-time compute

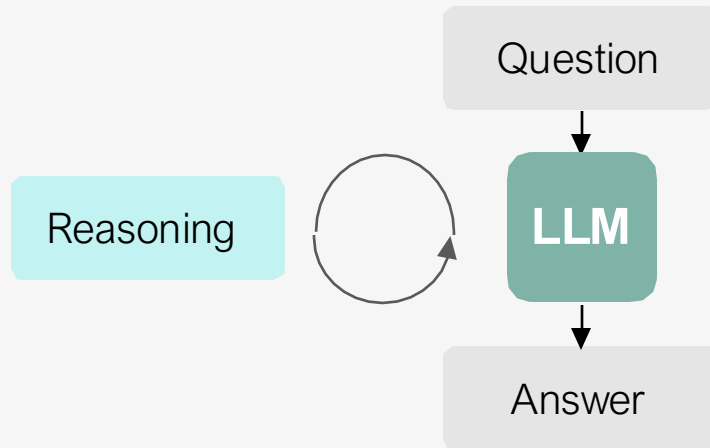
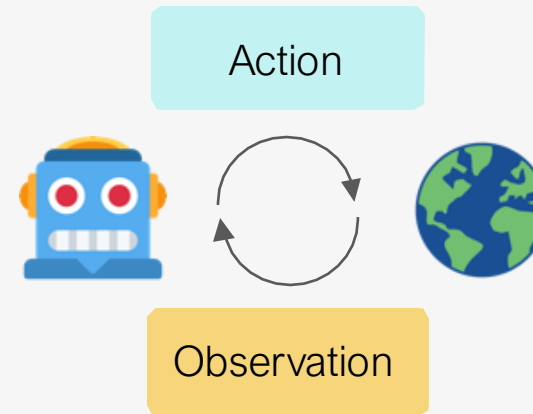
Lack of external knowledge and tools

RAG/Retrieval/Code/Tool use

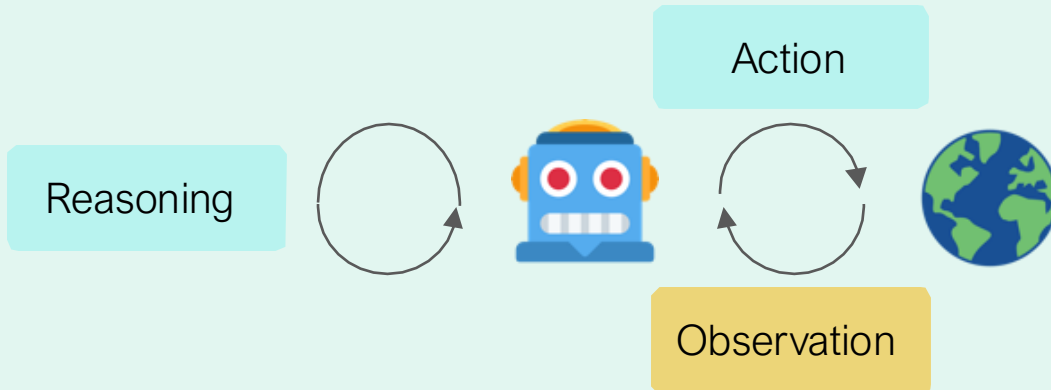


Flexible and general to
augment knowledge,
computation, feedback, etc.

Lack of reasoning

Reasoning (update internal belief)**Acting** (obtain external feedback)

ReAct: a new paradigm of agents that **reason and act**

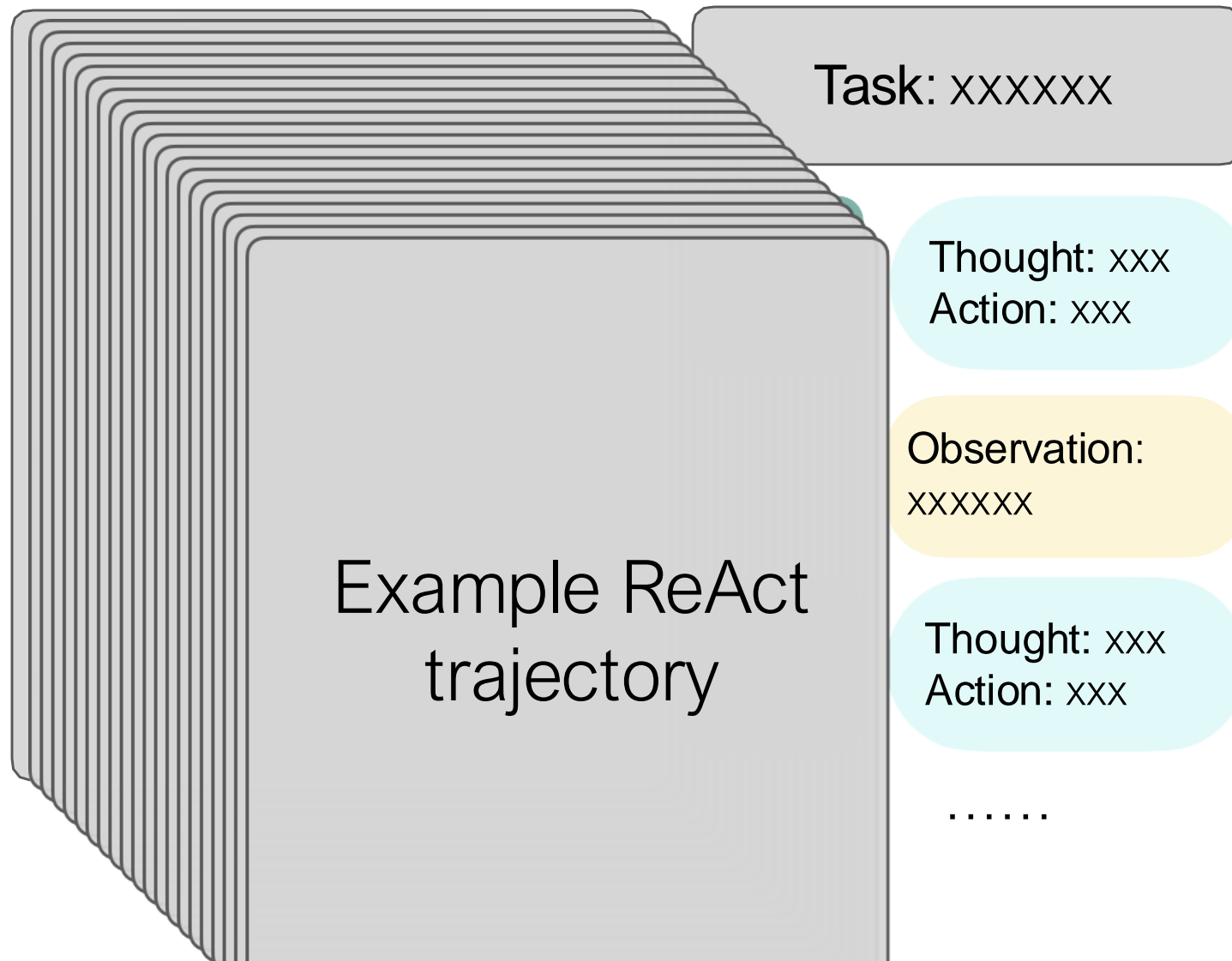
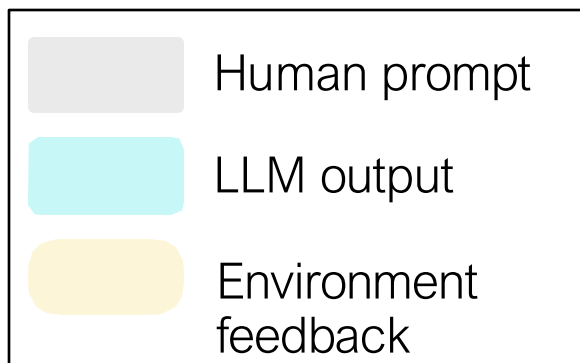


- **Synergy** of reasoning and acting
- **Simple** and intuitive to use
- **General** across domains

ReAct is simple and intuitive to use

ReAct supports

- One-shot prompting
- Few-shot prompting
- Fine-tuning



A zero-shot ReAct prompt

You are an agent that answers questions by using two actions:

- search[query]: Google search the query. You can also use it to calculate math.
- finish[answer]: return the answer

Your generation should have the following format:

Thought: to analyze the observation and inform the action.

Action: your action

Question: If I have seven trillion dollars today, can I buy Apple, Nvidia, and Microsoft? If not, how much more money do I need?



(GPT-4)

Thought: I need to find the current market capitalization of Apple, Nvidia, and Microsoft to understand how much it would cost to buy all of them.

Action: search["current market capitalization of Apple, Nvidia, and Microsoft"]


Observation: "Nvidia's market cap stood at \$1.832 trillion as of about 10 a.m. ET Wednesday, while Alphabet's was at \$1.814 trillion, bringing Nvidia behind just Microsoft (MSFT) at \$3.015 trillion and Apple (AAPL) at \$2.836 trillion."

Thought: Now that I have the market cap for all companies, let's add them together and see if \$7 trillion is enough.

Action: search["1.832 trillion + 3.015 trillion + 2.836 trillion"]

Observation: 7683000000000

Thought: The amount of money needed in addition to the \$7 trillion is \$683 billion.

Action: finish["You need an additional \$683 billion to be able to buy Apple, Nvidia, and Microsoft."] 



(Mar 2024)



(GPT-4)

Thought: I need to find the current market capitalization of Apple, Nvidia, and Microsoft to understand how much it would cost to buy all of them.

Action: search["current market capitalization of Apple, Nvidia, and Microsoft"]

Observation: nothing found!

(My adversarial observation)

Thought: The search did not give any results. I will need to search for each company's market capitalization individually.

Action: search["current market capitalization of Apple"]

Observation: 'price': 180.75

(My adversarial observation)

Thought: This is the current price of Apple's stock, not its market capitalization. The market cap is calculated by multiplying the share price by the number of outstanding shares. Thus, I need to find the total number of Apple's outstanding shares to calculate its market cap.

Action: search["total number of Apple's outstanding shares"]



Continuous to solve the task

Synergy: acting support reasoning, reasoning guides acting

Acting without Reasoning

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 dish sponge 2, and a spoon 1.

Act 3: Go to sinkbasin 1

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

Act 4: Take peppershaker 1 from sinkbasin 1

Obs 4: Nothing happens.

Act 5: Take peppershaker 1 from sinkbasin 1

Obs 5: Nothing happens.

.....

X

Cannot explore systematically or incorporate feedback

ReAct Enables Systematic Exploration

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

Act 1: **Think**[First I need to find a peppe 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

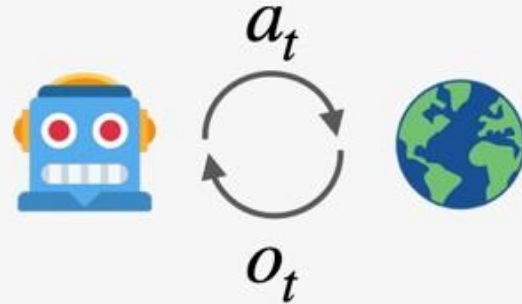
Obs 14: You put pepper shaker 1 in/on the drawer 1.



ReAct is general and effective

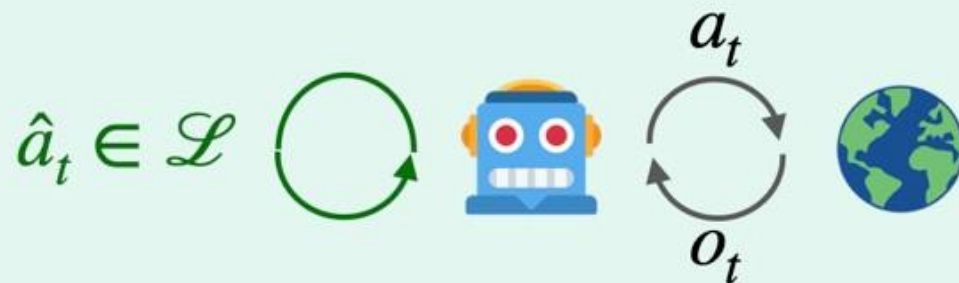
PaLM-540B	(NLP tasks)		(RL tasks)
	HotpotQA (QA)	FEVER (fact check)	ALFWorld (Text game)
Reason	29.4	56.3	N/A
Act	25.7	58.9	45
ReAct	35.1	64.6	71

Traditional agents: action space A defined by the environment



- **External feedback** o_t
- Agent context $c_t = (o_1, a_1, o_2, a_2, \dots, o_t)$
- Agent action $a_t \sim \pi(a | c_t) \in A$

ReAct: action space $\hat{A} = A \cup \mathcal{L}$ augmented by reasoning



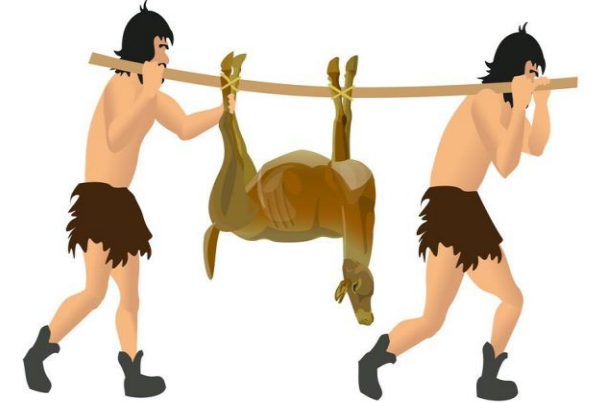
- $\hat{a}_t \in \mathcal{L}$ can be any language sequence
- Agent context $c_{t+1} = (c_t, \hat{a}_t, a_t, o_{t+1})$
- $\hat{a}_t \in \mathcal{L}$ only updates **internal context**

Reasoning agent: reasoning is an internal action for agents

Today's lecture

1. What are agents?
2. A brief history of LLM-powered agents in the context of LLMs and agents
3. Overview of LLM-powered agents
- 4. Tool use and learning**
5. Examples of agents in biology and medicine

Tool use and learning



**Human
intelligence**

**Small brain
capacity**

**Big brain
capacity**

Tool use

Collaborative work

**Artificial
intelligence**

Small models

Large models

**Autonomous
agents**

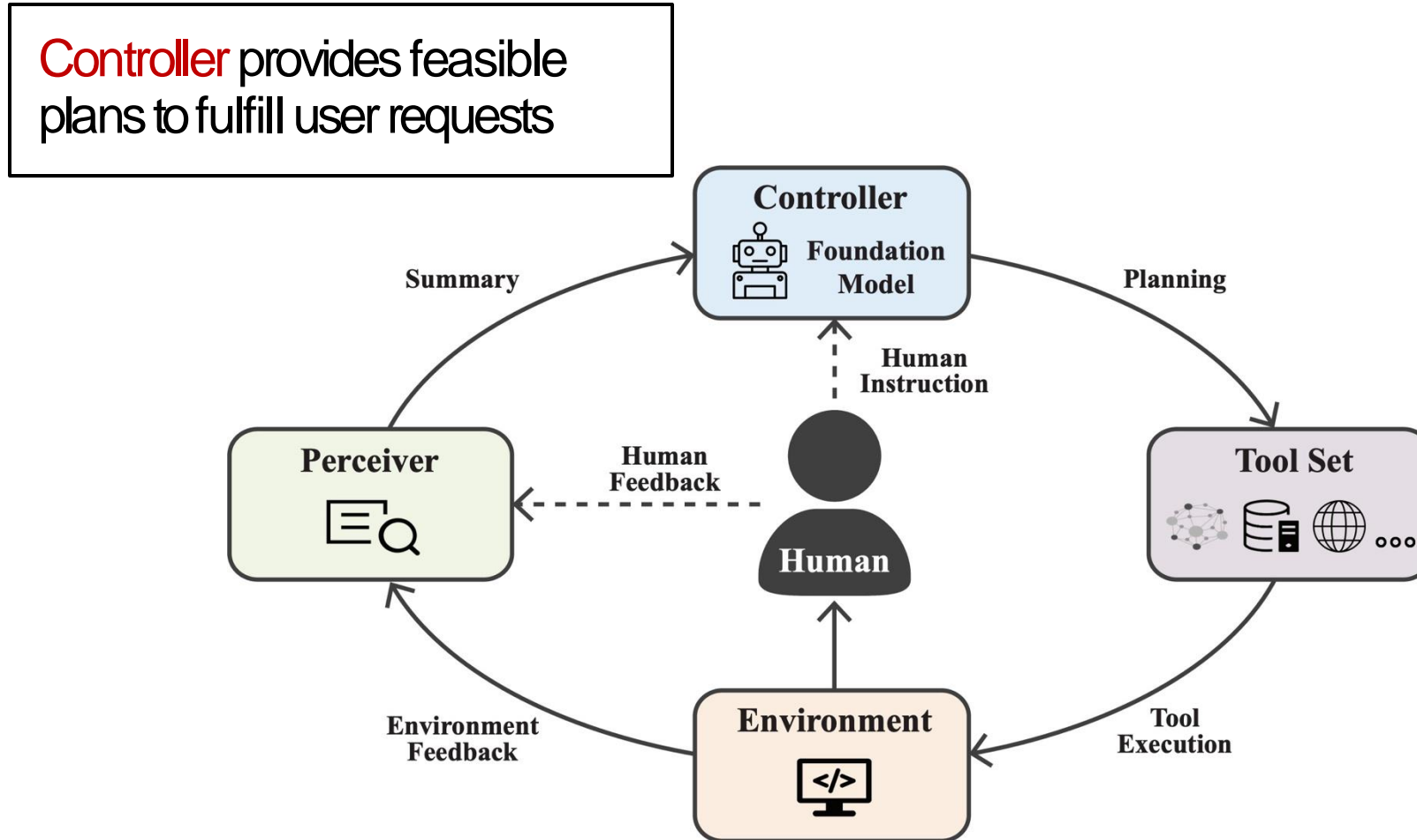
Multi-agent systems

Tool Intelligence

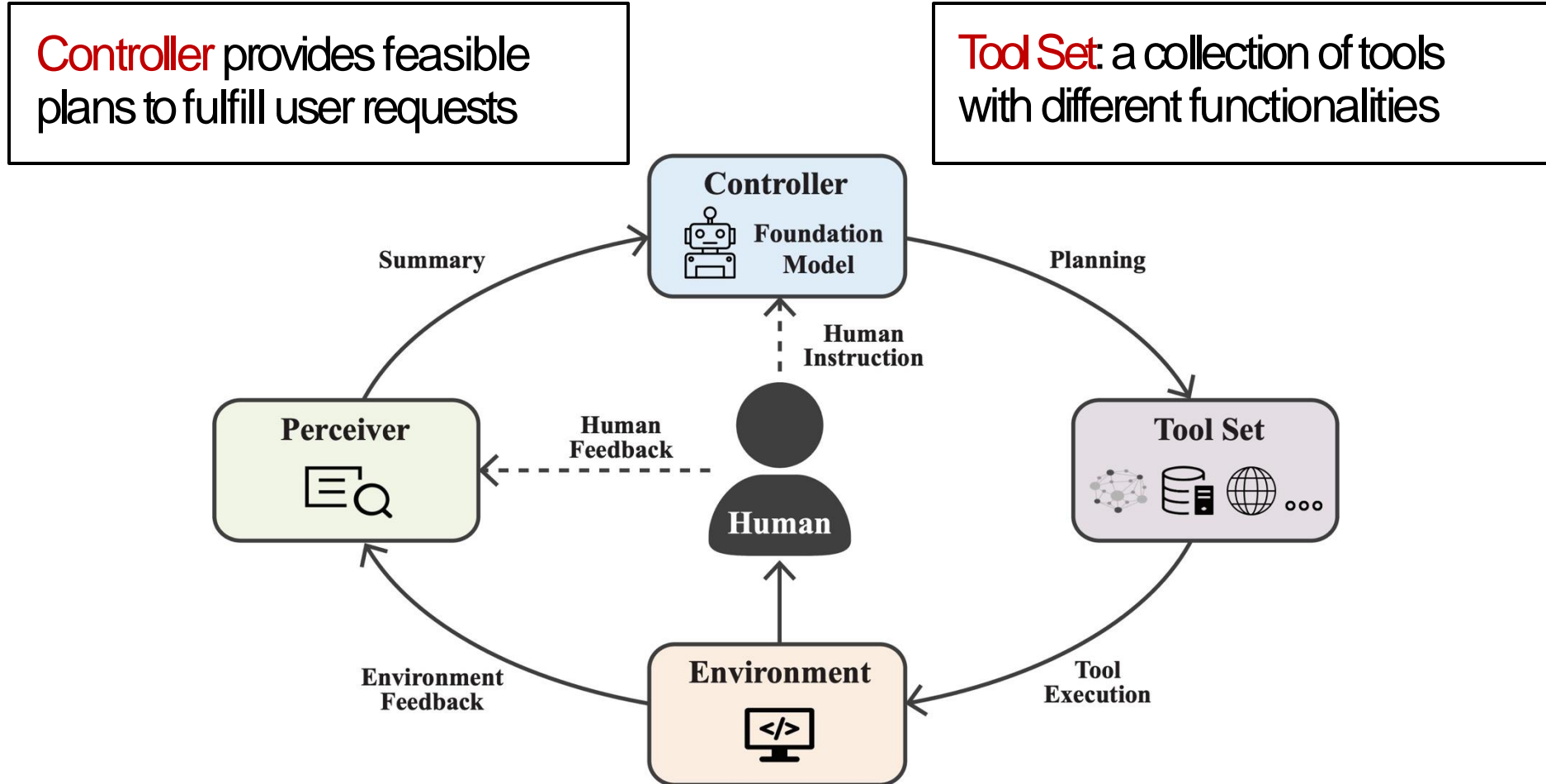
- Tools extends human capabilities in productivity, efficiency, and problem-solving
- Humans have been the **primary agents** in tool use throughout history
- Question: can **artificial intelligence** be as capable as humans in tool use?



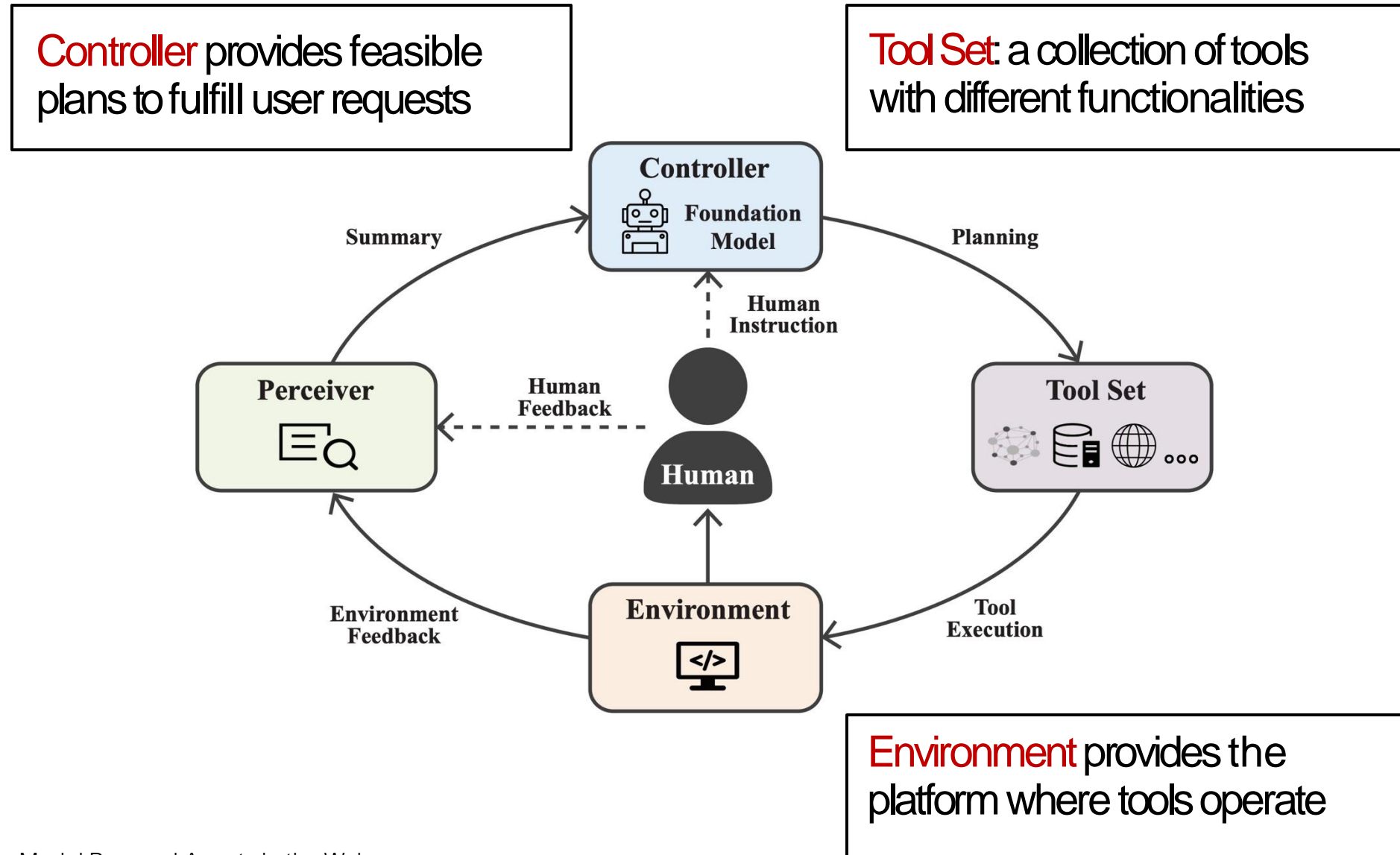
Framework for tool use in agents



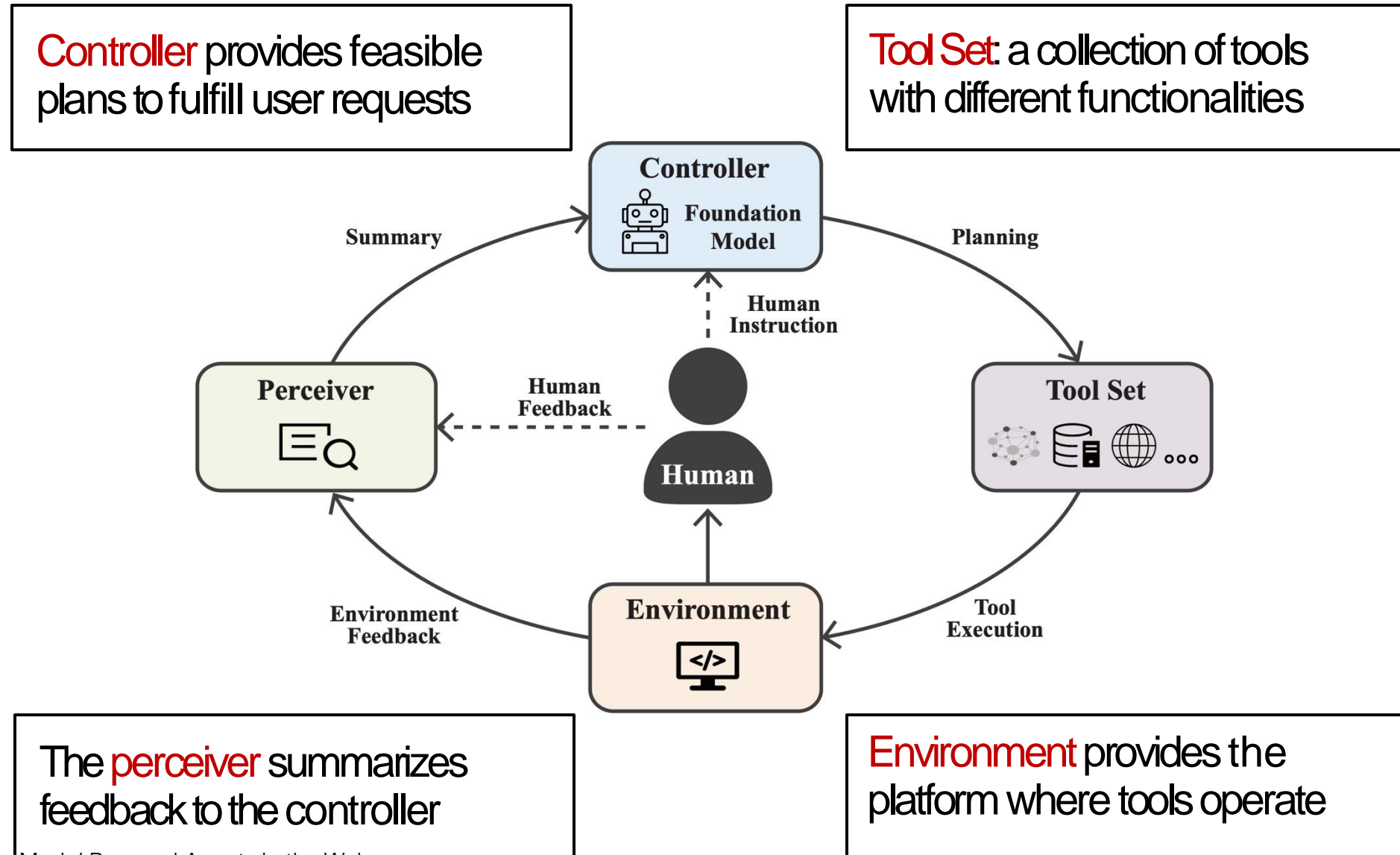
Framework for tool use in agents



Framework for tool use in agents



Framework for tool use in agents



Framework for tool use in agents

- Controller \mathcal{C} generates a plan a_t

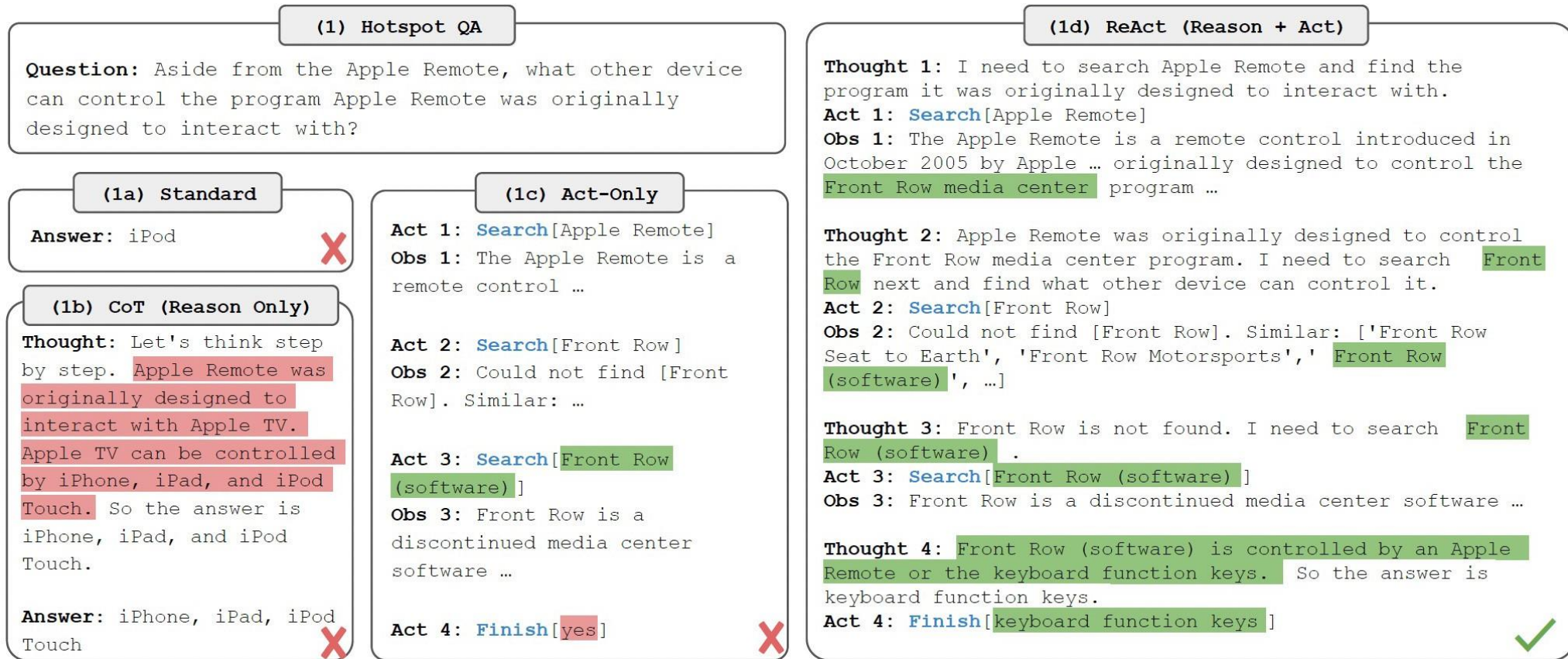
$$p_{\mathcal{C}}(a_t) = p_{\theta_{\mathcal{C}}}(a_t \mid \boxed{x_t}, \boxed{\mathcal{H}_t}, \boxed{q})$$

Feedback **History** **Instruction**

- **Problem**
 - **Planning:** divide the user query into sub-tasks
 - **Tool Use:** use the appropriate tool to solve sub-task
 - **Memory:** manage the working history
 - **Profile:** manage the user preference

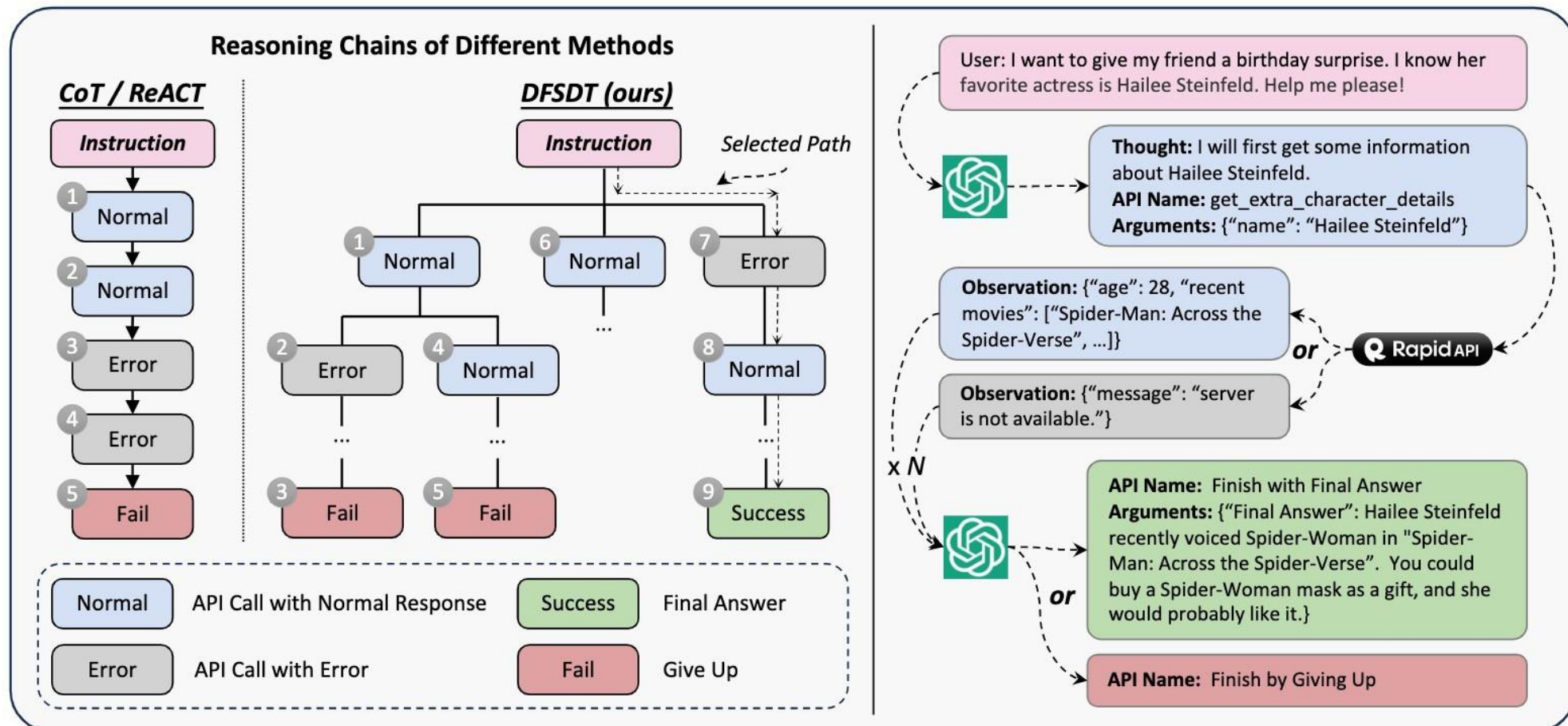
Planning with feedback

- ReAct



Planning with feedback

- DFSDT - Depth-first search-based decision tree method for reasoning



User: I want to give my friend a birthday surprise. I know her favorite actress is Hailee Steinfeld. Help me please!

Thought: I will first get some information about Hailee Steinfeld.
API Name: get_extra_character_details
Arguments: {"name": "Hailee Steinfeld"}

Observation: {"age": 28, "recent movies": ["Spider-Man: Across the Spider-Verse", ...]}

Observation: {"message": "server is not available."}

API Name: Finish with Final Answer
Arguments: {"Final Answer": "Hailee Steinfeld recently voiced Spider-Woman in "Spider-Man: Across the Spider-Verse". You could buy a Spider-Woman mask as a gift, and she would probably like it."}

API Name: Finish by Giving Up

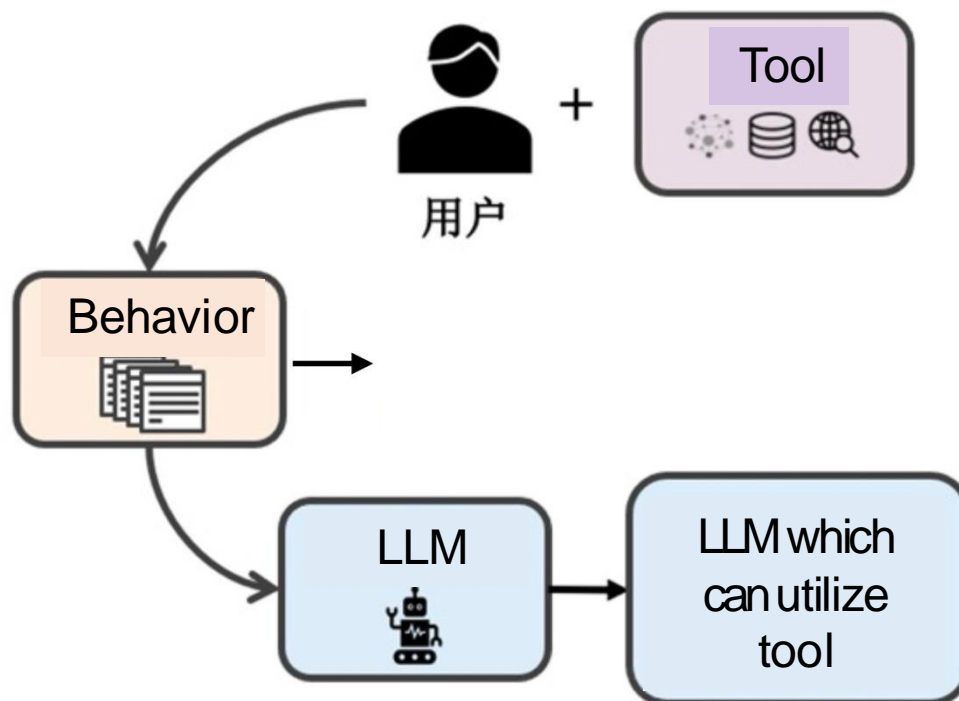
x N

Learning to use tools


- Imitation Learning

- By recording data on human tool usage behaviors, large models mimic human actions to learn about tools

- The simplest and most direct method of tool learning.



WebGPT

- Supervised Learning  OpenAI
 - Clone human behavior to use search engines
 - Supervised fine-tuning + reinforcement learning
 - Only need 6,000 annotated data points

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.

↑↑ [1] Gifts From Crows | Outside My Window (www.birdsoutsidemymywindow.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.

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

Done quoting! Write an answer

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

◆Quotes
From Gifts From Crows | Outside My Window (www.birdsoutsidemymywindow.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.birdsoutsidemymywindow.org
Quote
Back

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

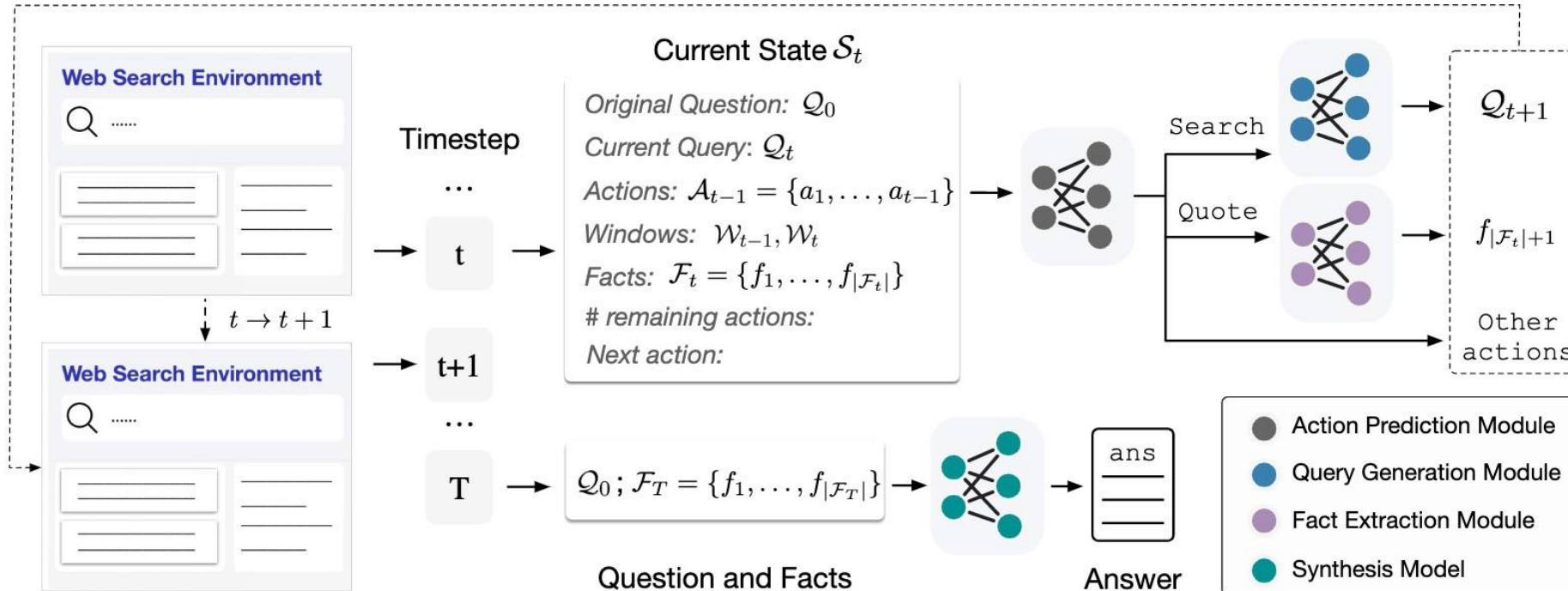
◆Scrollbar: 0 - 11
◆Text
{0}How to Make Friends With Crows - PetHelpful{pethelpful.com}
If you did this a few times, your crows would learn your new place, but as I said, I'm not sure if they will follow or visit you there since it's probably not in their territory. The other option is simply to make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

{1}Gifts From Crows | Outside My Window{www.birdsoutsidemymywindow.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

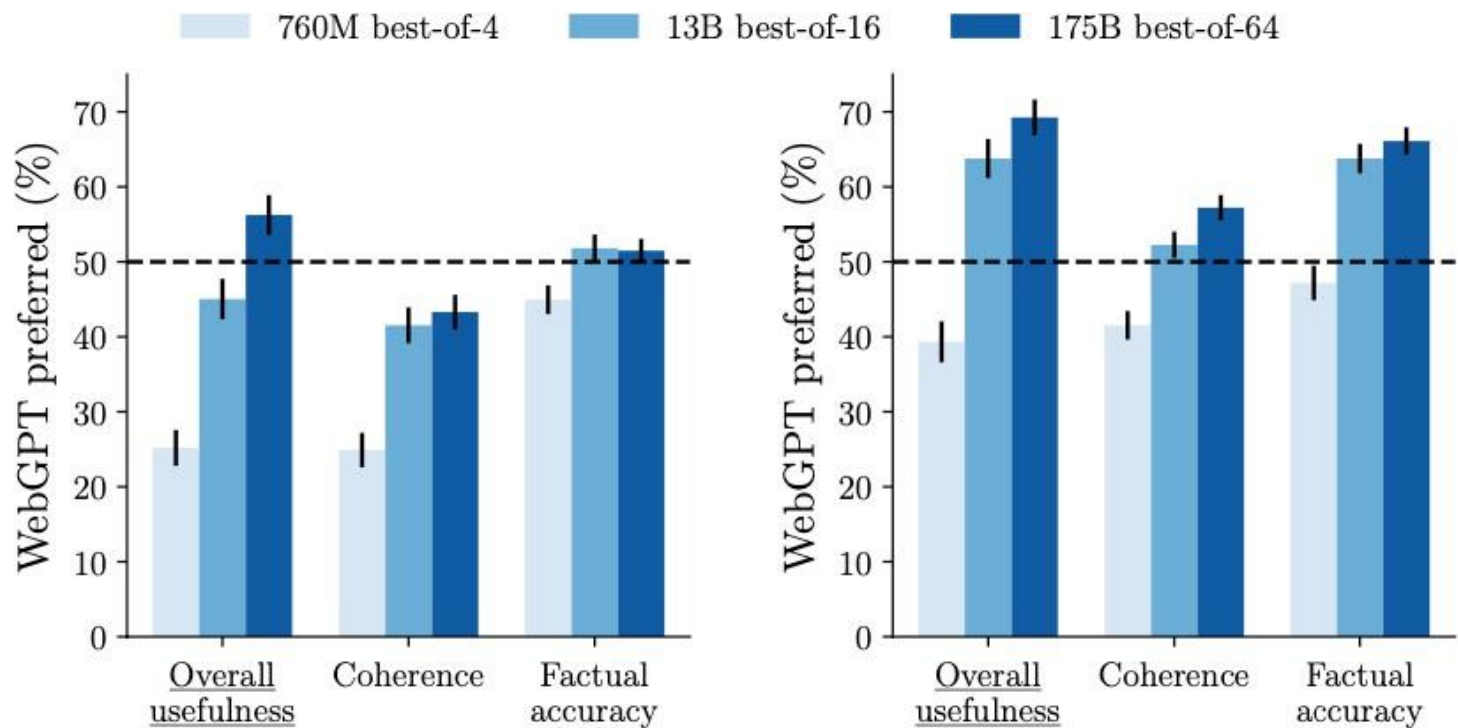
WebGPT

- At each step, the **search model** executes actions to collect supporting facts, which are sent to the **synthesis model** for answer generation



WebGPT

- Supervised Learning  OpenAI
 - Excellent performance in long-form QA, even surpassing human experts

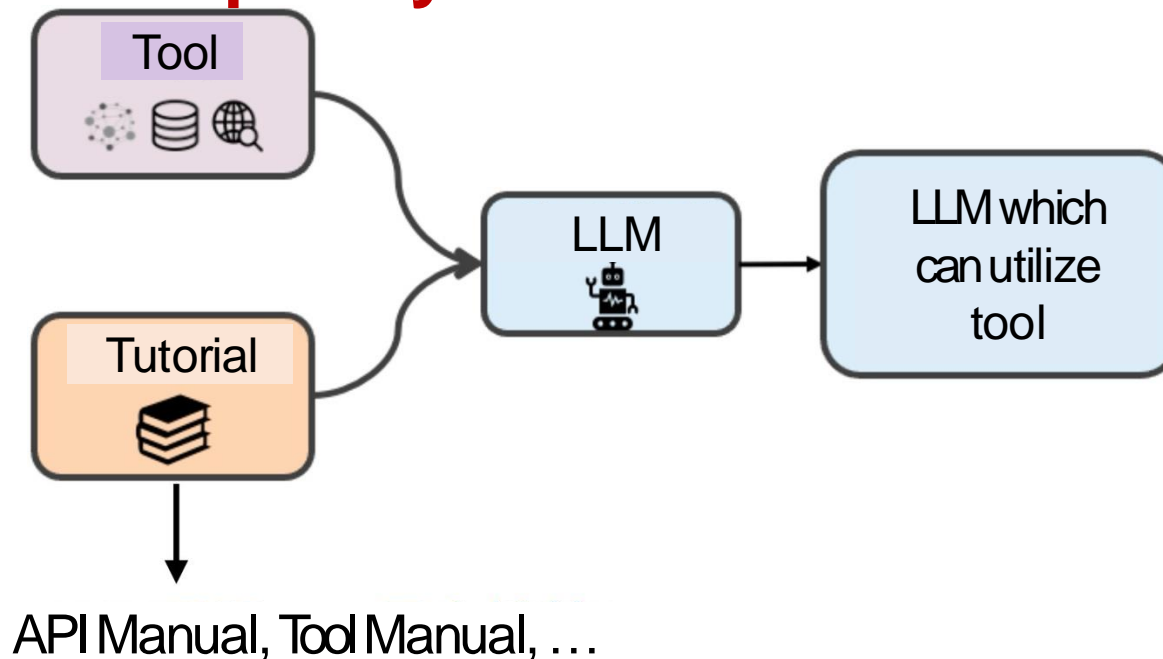


(a) WebGPT vs. human demonstrations.

(b) WebGPT vs. ELI5 reference answers.

Learning to use tools

- Tutorial Learning
 - By having the model read tool manuals (tutorials), it understands the functions of the tools and how to invoke them
- Almost exclusively, large models from the OpenAI series (such as ChatGPT, GPT-4) possess a high **zero-shot capability** to understand tool manuals.



Learning to use tools

- Describe the functionality; In-context with example(s).

Zero-shot Prompting: Here we provide a tool (API) "forecast_weather(city:str, N:int)", which could forecast the weather about a city on a specific date (after N days from today). The returned information covers "temperature", "wind", and "precipitation".

Please write codes using this tool to answer the following question: "What's the average temperature in Beijing next week?"

Few-shot Prompting: We provide some examples for using a tool. Here is a tool for you to answer question:

Question: "What's the temperature in Shanghai tomorrow?"

```
return forecast_weather("Shanghai", 1) ["temperature"]
```

Question: "Will it rain in London in next two days?"

```
for i in range(2):
    if forecast_weather("London", i+1) ["precipitation"] > 0:
        return True
return False
```

Question: "What's the average temperature in San Francisco next week?"

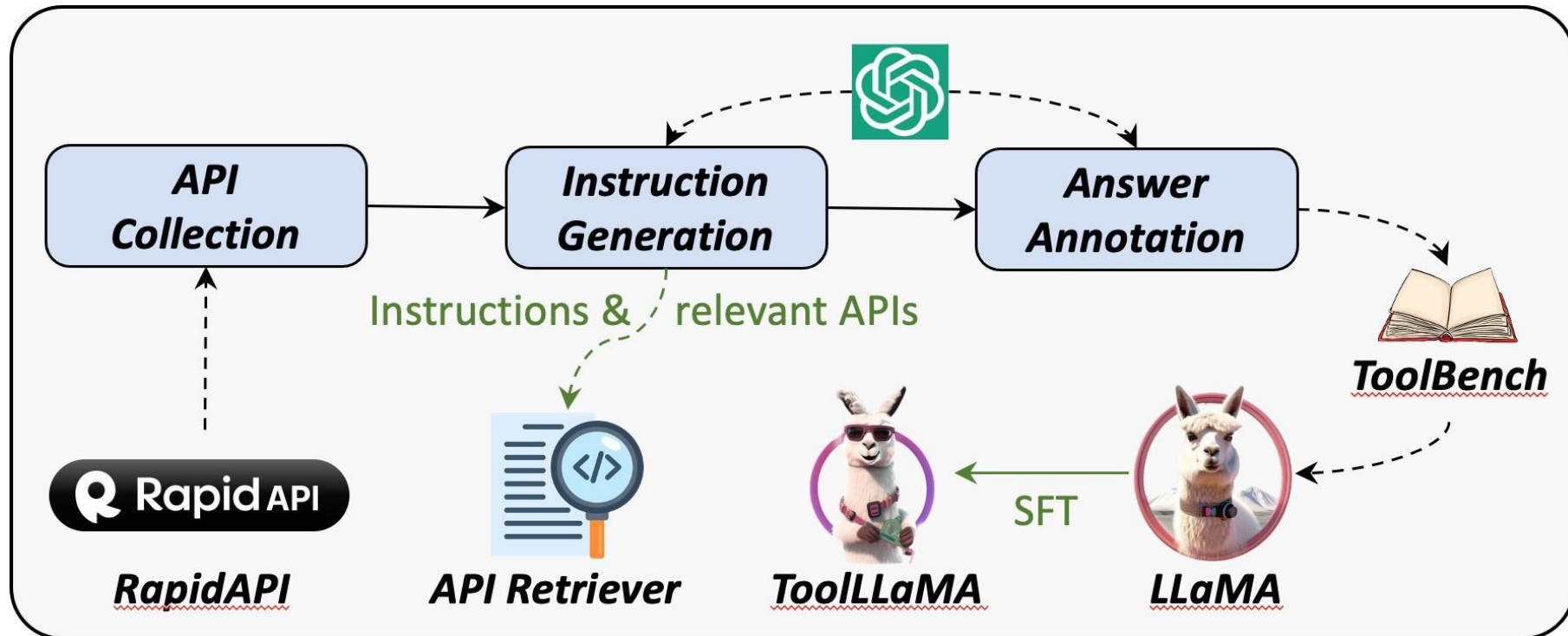
ToolBench

- Highlights:
 - Over 16,000 real APIs (collected from RapidAPI)
 - Supports single and multi-tool invocation
 - Complex multi-step reasoning tasks

Resource	ToolBench (this work)	APIBench (Patil et al., 2023)	API-Bank (Li et al., 2023a)	ToolAlpaca (Tang et al., 2023)	T-Bench (Xu et al., 2023b)
Real-world API?	✓	✗	✓	✗	✓
Real API Response?	✓	✗	✓	✗	✓
Multi-tool Scenario?	✓	✗	✗	✗	✗
API Retrieval?	✓	✓	✗	✗	✗
Multi-step Reasoning?	✓	✗	✓	✓	✓
Number of tools	3451	3	53	400	8
Number of APIs	16464	1645	53	400	232
Number of Instances	12657	17002	274	3938	2746
Number of Real API Calls	37204	0	568	0	0
Avg. Reasoning Traces	4.1	1.0	2.1	1.0	5.9

ToolBench construction

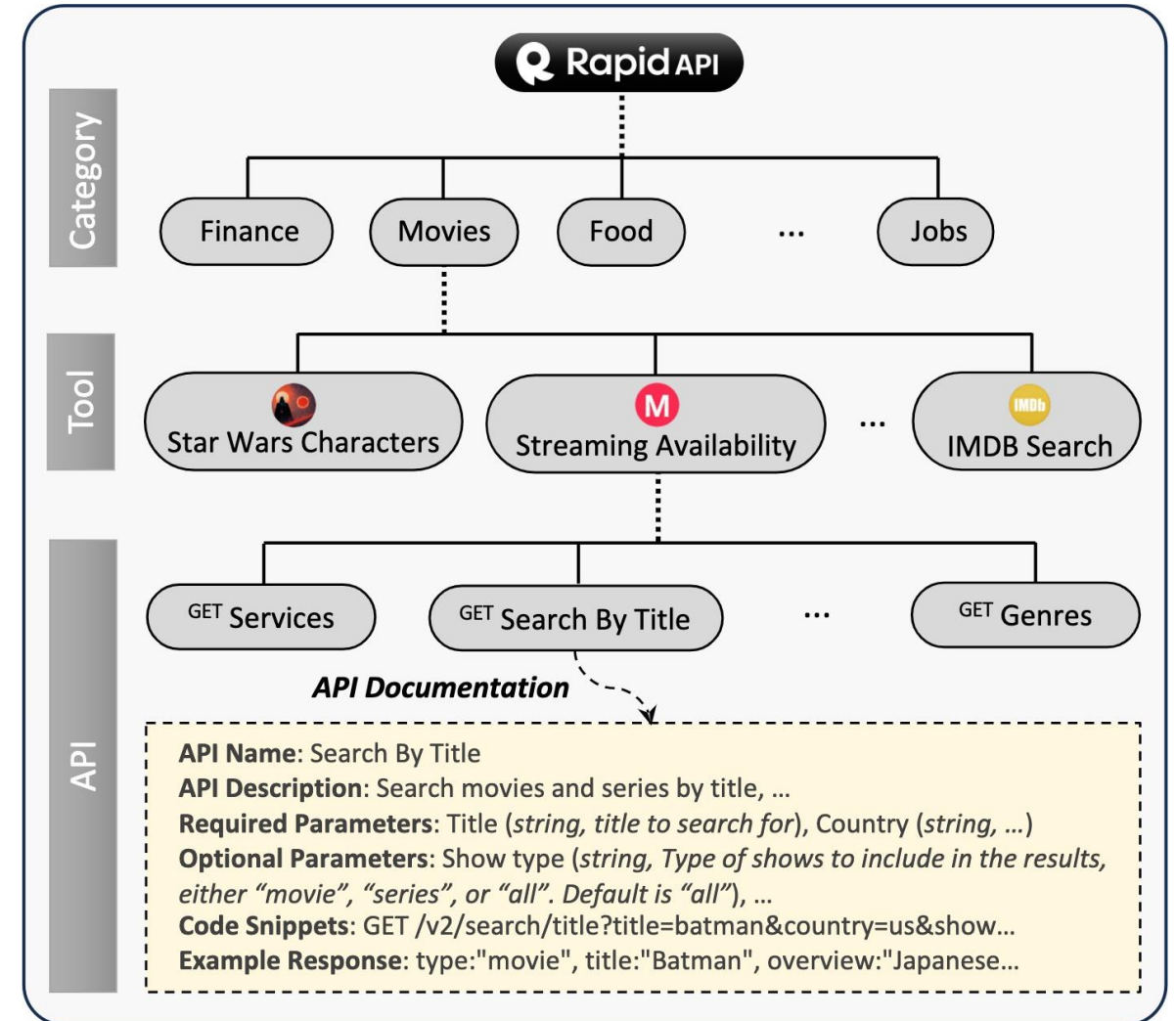
- API Collection
- Instruction Generation
- Answer Annotation



ToolBench construction

- API Collection

- RapidAPI Hub: <https://rapidapi.com/hub>
- Filter over 16,000 high-quality APIs from more than 50,000 APIs
- Include 49 categories



ToolBench construction

- Instruction generation

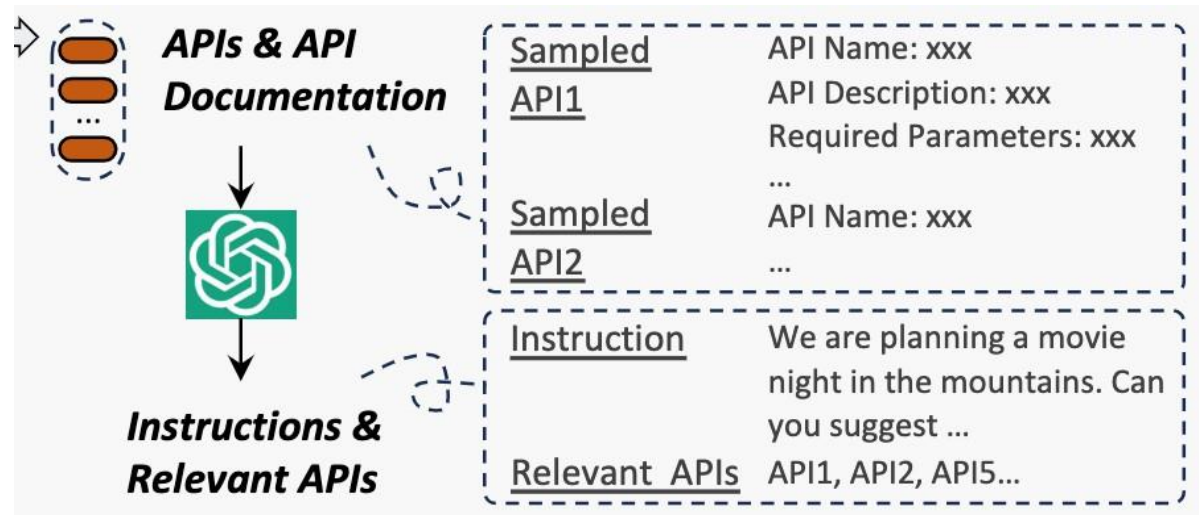
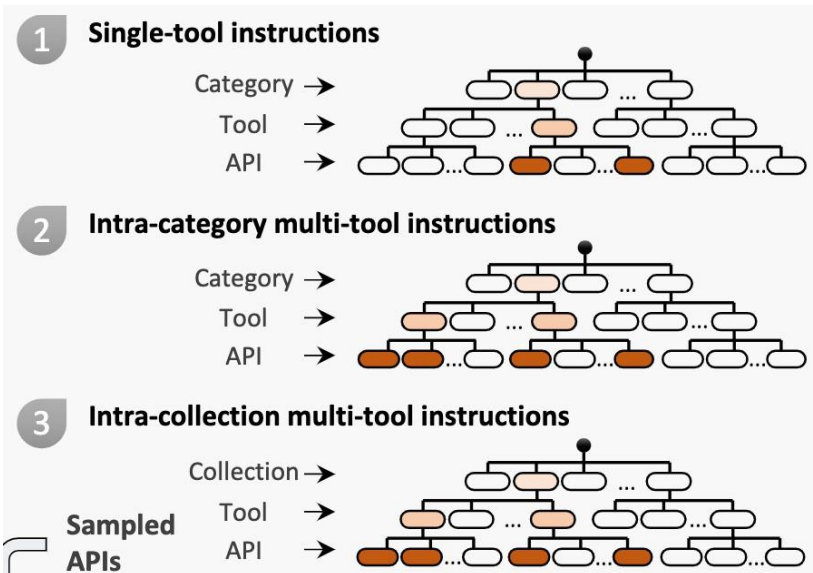
- Single Tool + Multi-Tool

- (1) Sample a collection of APIs: $\mathbb{S}_N^{\text{sub}} = \{\text{API}_1, \dots, \text{API}_N\}$

- (2) ChatGPT automatically generate instructions that may require calling one or more APIs in the collection:

$$\text{ChatGPT} \left(\left\{ [\mathbb{S}_1^{\text{rel}}, \text{Inst}_1], \dots, [\mathbb{S}_{N'}^{\text{rel}}, \text{Inst}_{N'}] \right\} \mid \text{API}_1, \dots, \text{API}_N, \text{seed}_1, \dots, \text{seed}_3 \right).$$

$\{\text{API}_1, \dots, \text{API}_N\} \in \mathbb{S}_{\text{API}}$
 $\{\text{seed}_1, \dots, \text{seed}_3\} \in \mathbb{S}_{\text{seed}}$



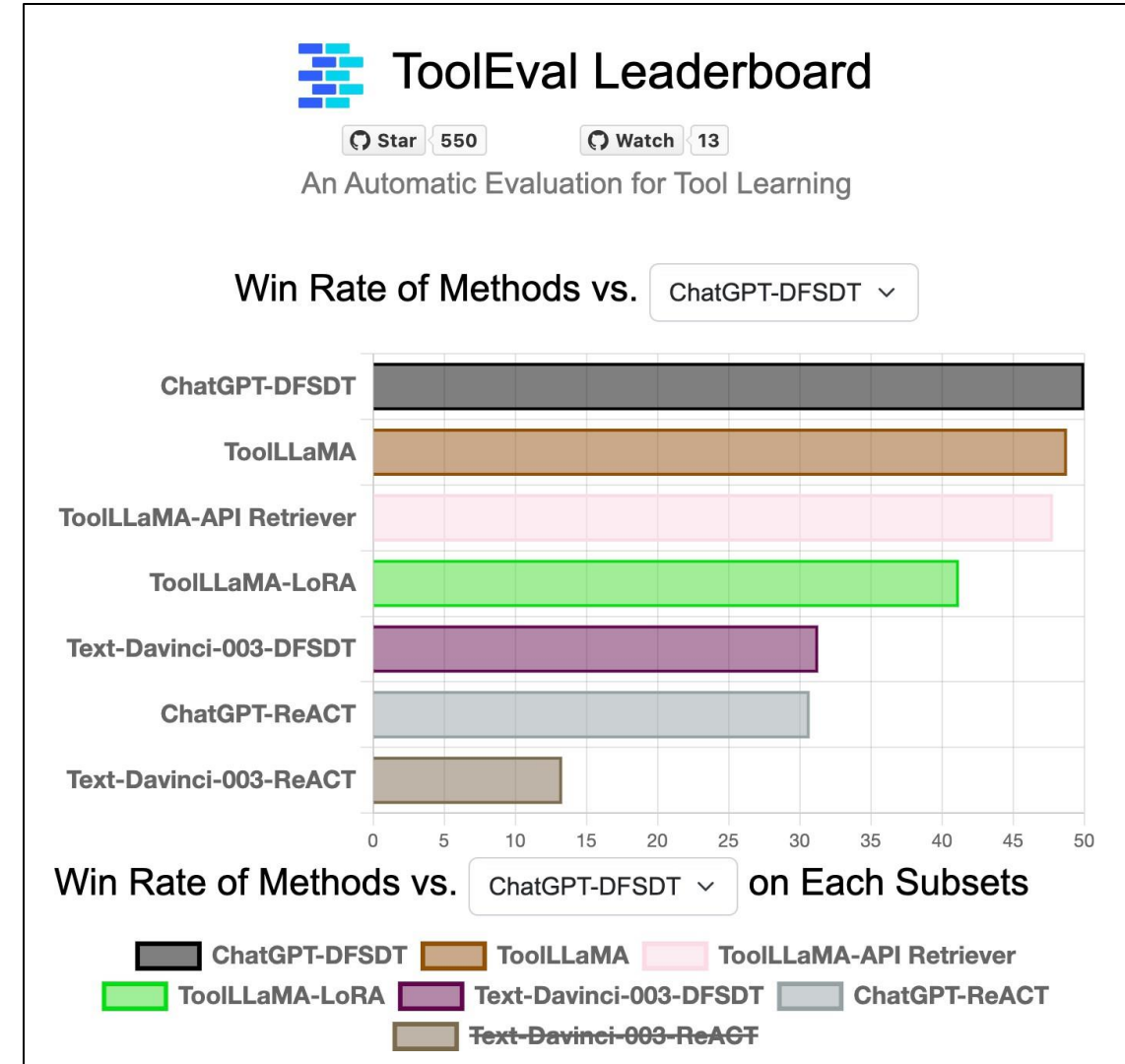
ToolBench construction

- Answer Annotation
 - gpt-3.5-turbo-16k: feature of function call
- Issues with ReACT
 - Error Propagation: An error in a single step annotation can render the entire action sequence unusable
 - Limited Exploration: ReACT can only sample one sequence from the infinite action sequence space based on the LM's probabilities
- DFSDT: Dynamically extends the TOT to the tool learning scenario

Method	Single-tool (I1)	Category (I2)	Collection (I3)	Average
ReACT	43.98	23.62	20.42	29.34
ReACT@N	50.80	36.14	32.87	39.94
DFSDT	54.10	47.35	44.80	48.75

ToolEval

- Automatic evaluation framework based on ChatGPT
- Two metrics:
 - **Success rate:** The proportion of commands successfully completed within a limited number of API calls
 - **Preference:** Comparison of quality/usefulness between two answers, i.e., which one is better?
- Highly consistent with human experts (~80%)



ToolLLaMA

- Demonstrate exceptionally high generalizability to OOD commands and APIs, significantly outperforming ChatGPT+ReACT

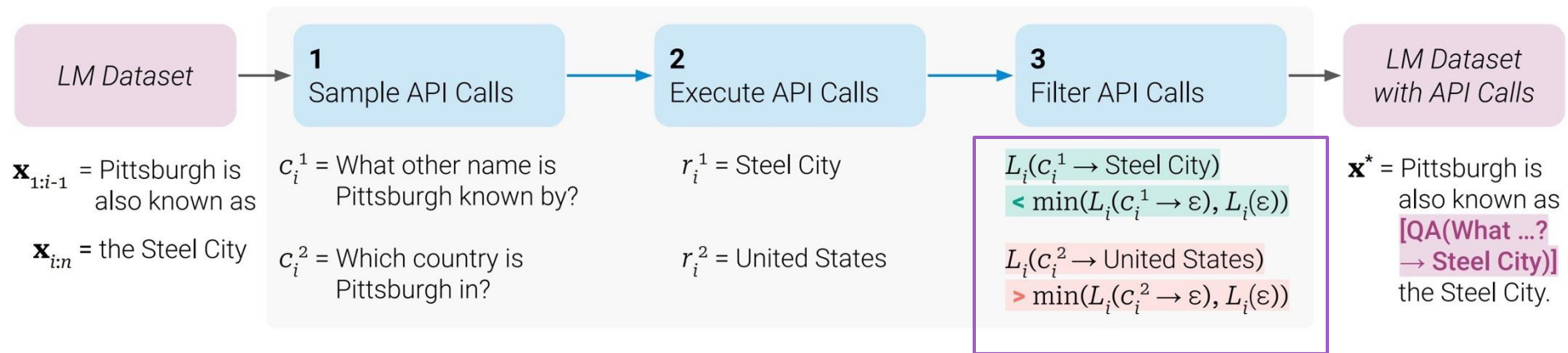
Model	I1-Inst.		I1-Tool		I1-Cat.		I2-Inst.		I2-Cat.		I3-Inst.		Average	
	Pass	Win	Pass	Win	Pass	Win	Pass	Win	Pass	Win	Pass	Win	Pass	Win
ChatGPT-ReACT	56.0	-	62.0	-	66.0	-	28.0	-	22.0	-	30.0	-	44.0	-
Vicuna (ReACT & DFSDT)	0.0	-	0.0	-	0.0	-	0.0	-	0.0	-	0.0	-	0.0	-
Alpaca (ReACT & DFSDT)	0.0	-	0.0	-	0.0	-	0.0	-	0.0	-	0.0	-	0.0	-
Text-Davinci-003-DFSDT	53.0	46.0	58.0	38.0	61.0	39.0	38.0	46.0	38.0	45.0	39.0	48.0	47.8	43.7
ChatGPT-DFSDT	78.0	68.0	84.0	59.0	89.0	57.0	51.0	78.0	58.0	<u>77.0</u>	57.0	77.0	69.6	69.3
ToolLLaMA-DFSDT	<u>68.0</u>	68.0	<u>80.0</u>	59.0	<u>75.0</u>	<u>56.0</u>	<u>47.0</u>	<u>75.0</u>	<u>56.0</u>	80.0	<u>40.0</u>	<u>72.0</u>	<u>61.0</u>	<u>68.3</u>

- DFSDT >> ReACT

Method	Single-tool (I1)	Category (I2)	Collection (I3)	Average
ReACT	43.98	23.62	20.42	29.34
ReACT@N	50.80	36.14	32.87	39.94
DFSDT	54.10	47.35	44.80	48.75

Toolformer

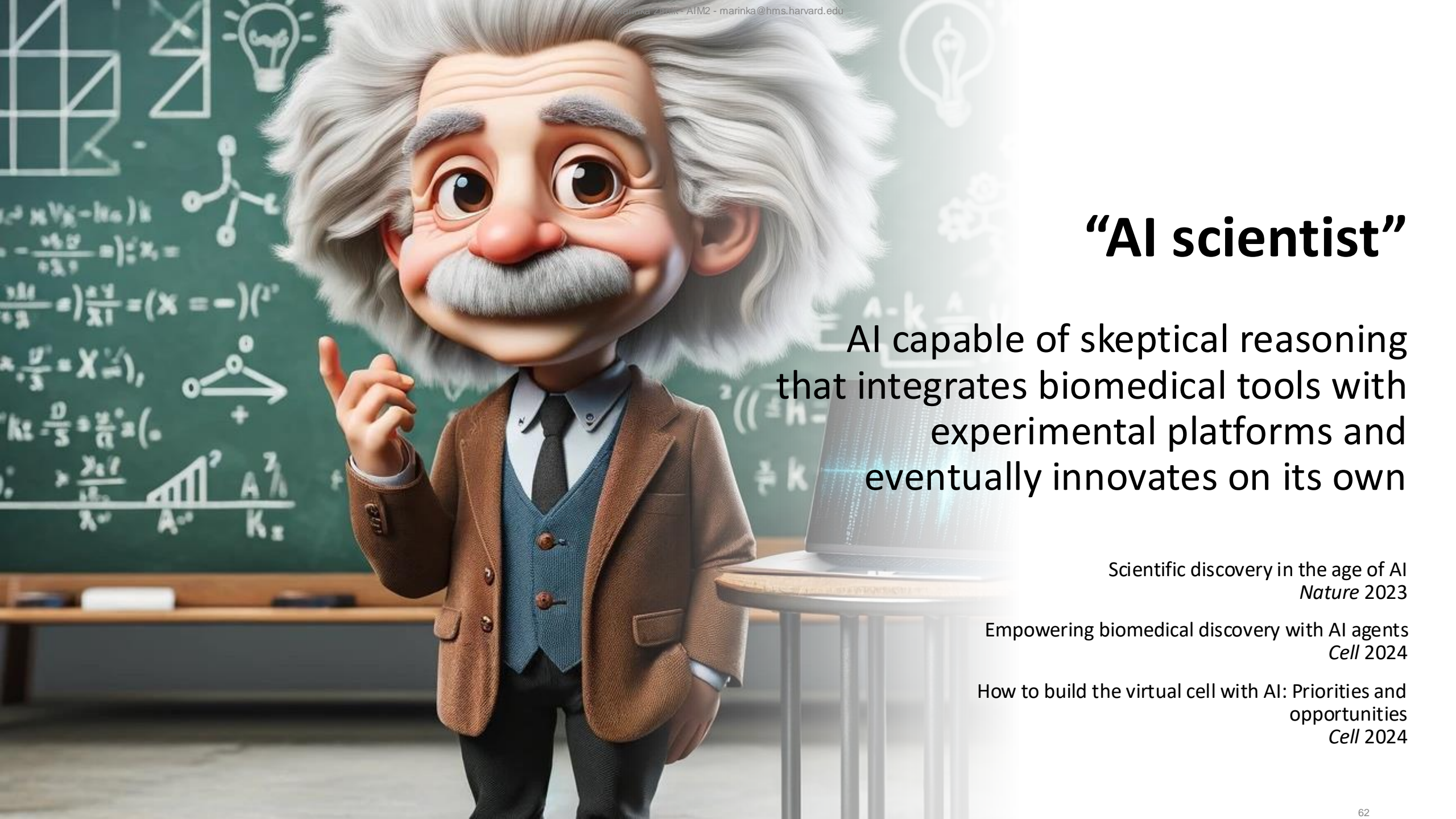
- Self-supervised tool learning
 - Pre-defined tool APIs
 - Encourage models to call and execute tool APIs
 - Design self-supervised loss to see if the tool execution can help language modeling



If the tool execution reduces LM loss,
save the instances as training data

Today's lecture

1. What are agents?
2. A brief history of LLM-powered agents in the context of LLMs and agents
3. Overview of LLM-powered agents
4. Tool use and learning
- 5. Examples of agents in biology and medicine**



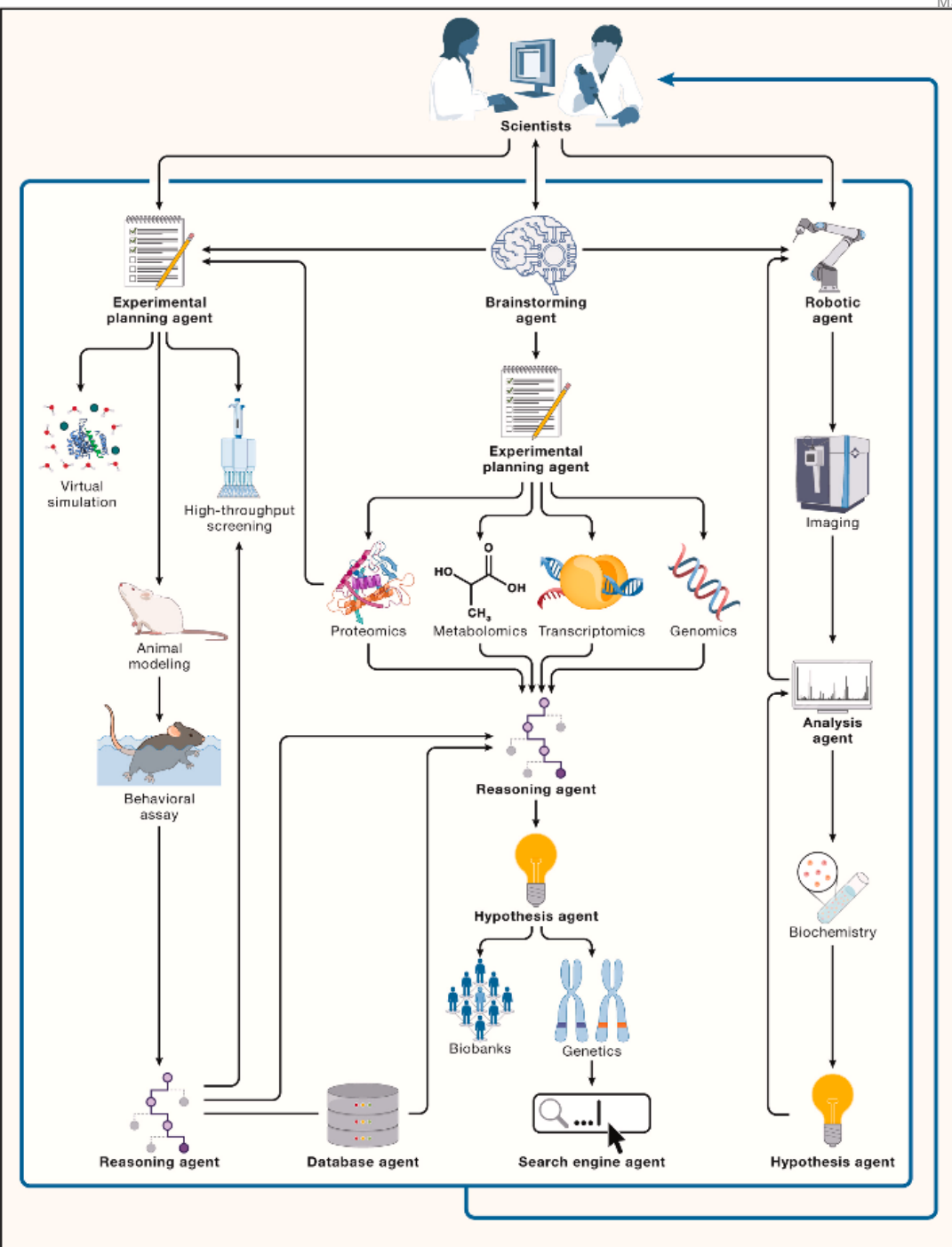
“AI scientist”

AI capable of skeptical reasoning that integrates biomedical tools with experimental platforms and eventually innovates on its own

Scientific discovery in the age of AI
Nature 2023

Empowering biomedical discovery with AI agents
Cell 2024

How to build the virtual cell with AI: Priorities and opportunities
Cell 2024

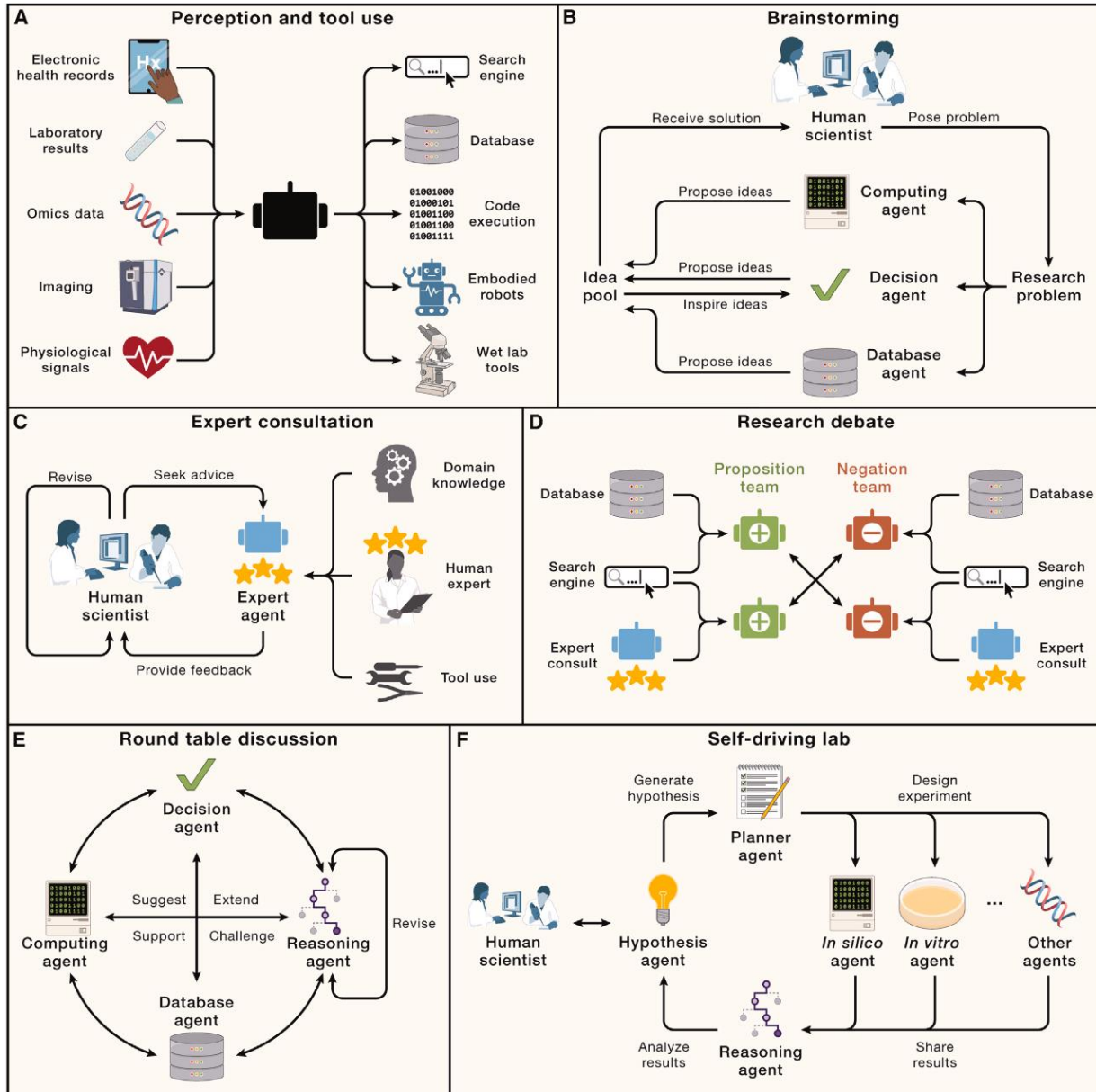


Empowering discovery with AI agents

- Integrate human creativity and AI capabilities
- Coordinate tasks like hypothesis creation, data analysis, and experiment design
- Use reasoning, interaction, and memory for reflective learning
- Combine tools like LLMs, FMs and ML models, and experimental platforms
- Collaborate with humans and other agents in modular workflows
- Vision: AI systems as collaborative partners in scientific discovery

Databases and search engines

- ◆ **October 1990**
BLAST
- **April 1998**
PageRank
- ◆ **January 2002**
Machine learning for DLBC outcome prediction
- ◆ **March 2002**
Patient outcome prediction via gene expression
- ◆ **April 2005**
Automated derivation of cellular signaling influences
- ◆ **November 2012**
Genetic variation map from 1,092 genomes
- **December 2012**
AlexNet
- ◆ **August 2015**
DeepBind
- ◆ **February 2016**
Genotype to phenotype via cell hierarchy



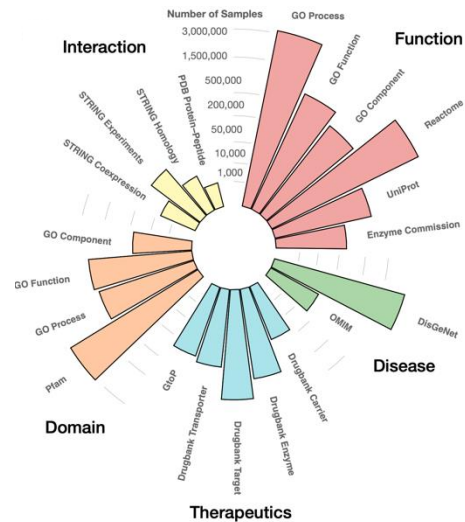
AI agent configurations

- **Perception and tool use** integrate multimodal data and utilize specialized tools to expand research capabilities
- **Brainstorming agents** propose diverse research ideas
- **Expert consultation agents** leverage domain-specific expertise to refine ideas and validate findings
- **Debate and roundtable agents** critically evaluate hypotheses
- **Self-driving lab agents** optimize end-to-end research workflows

Agentic AI: From molecules to therapies

Molecules

Knowledge-grounded AI agent for expert-level biological reasoning



Cells

Multi-agent AI for single-cell powered drug discovery



Find 5 gene candidates in goblet cells involved in non small cell lung cancer

The user are requested the top 5 genes in goblet cell involved in non-small cell lung carcinoma, **EFO: 0003060**.

The following genes in goblet cells are involved in non-small cell lung cancer:

ERBB, ALK, KRAS, BRAF, MET

References and Code supported the analysis can be found in:

Patients

AI agent with a massive number of tools for precision therapy



We have models that predict **protein structure** from **sequence** and achieve **atom-level accuracy**

Structure for each protein

Predicting **protein function** from sequence, structure **is an open challenge**

Function for each protein

20% of human proteins lack well-defined functions, and even well-annotated proteins often miss functional insights across various biological contexts and disease states

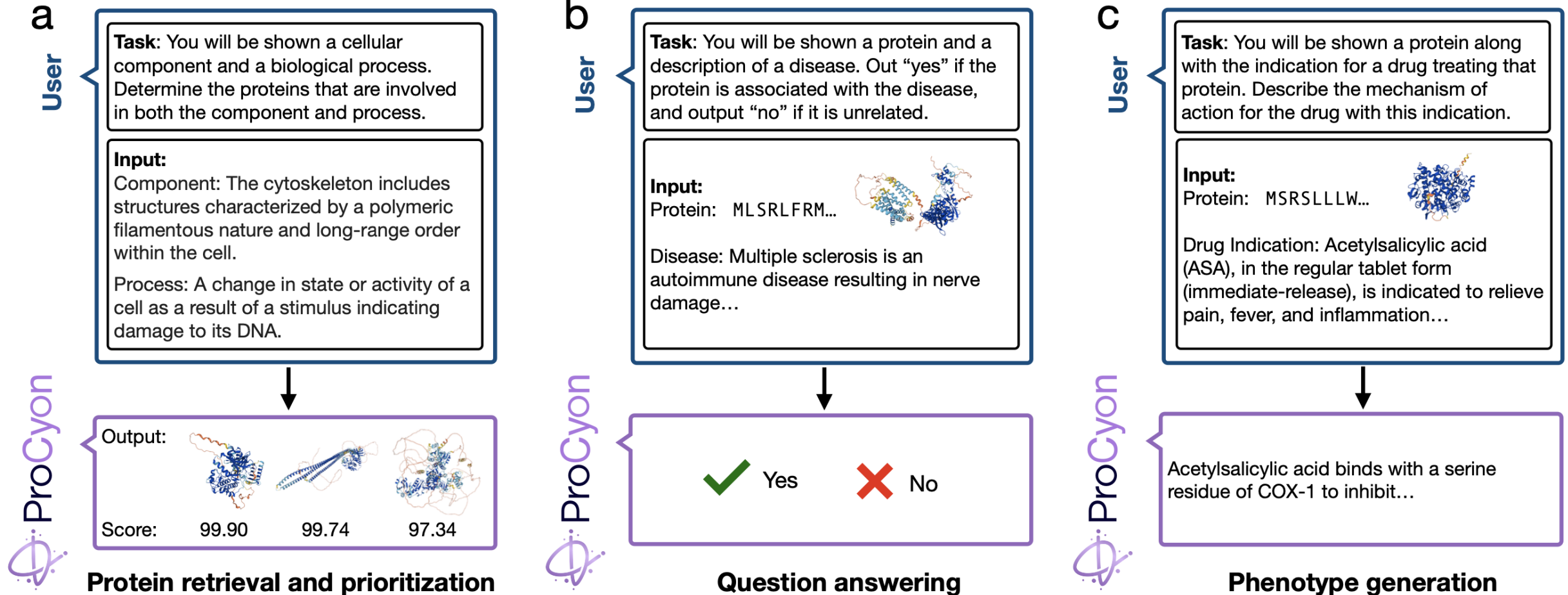
40% of human proteins are missing context-specific functional insights

95% of publications focus on only 5,000 human proteins

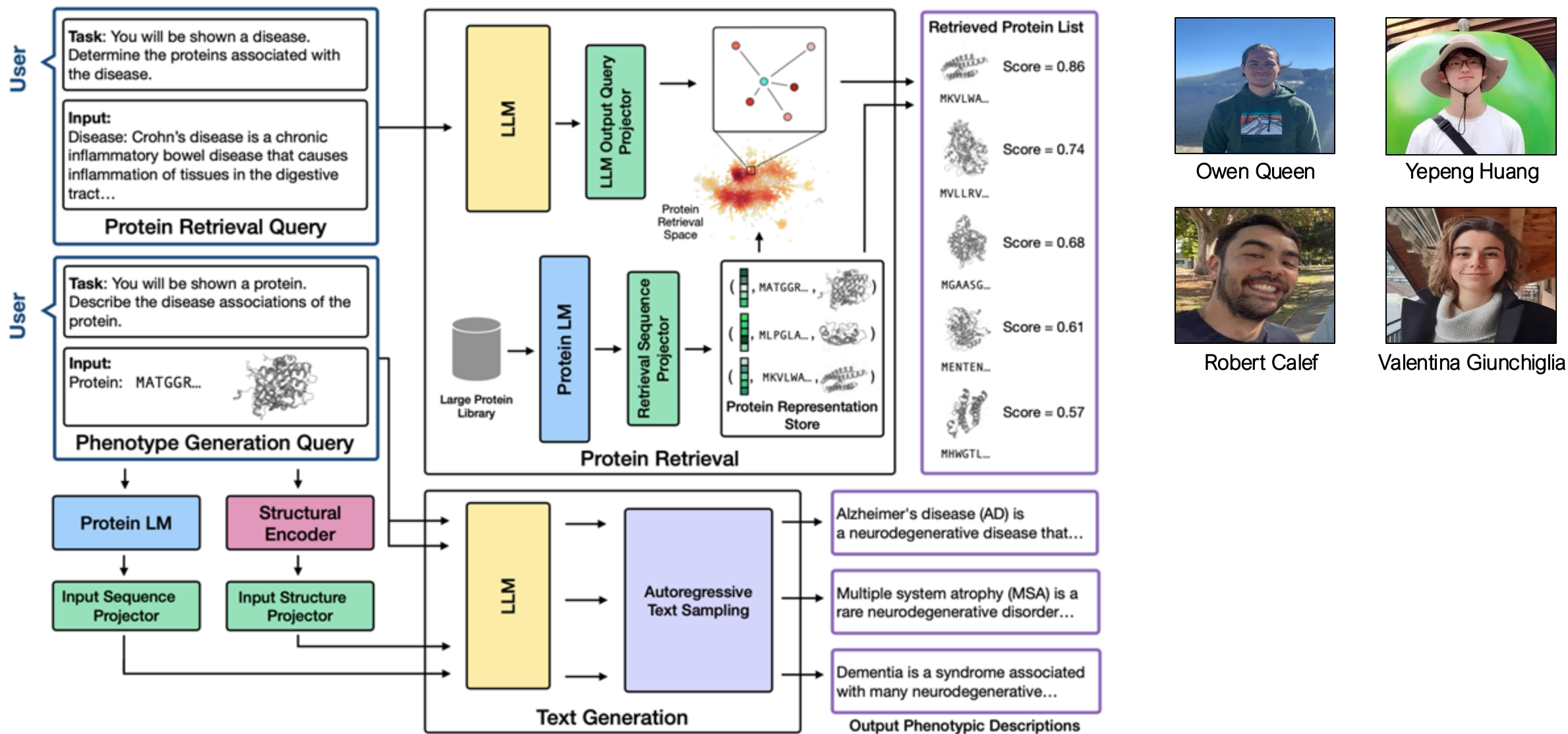


zitniklab.hms.harvard.edu/ProCyon

ProCyon: Foundation model for protein function

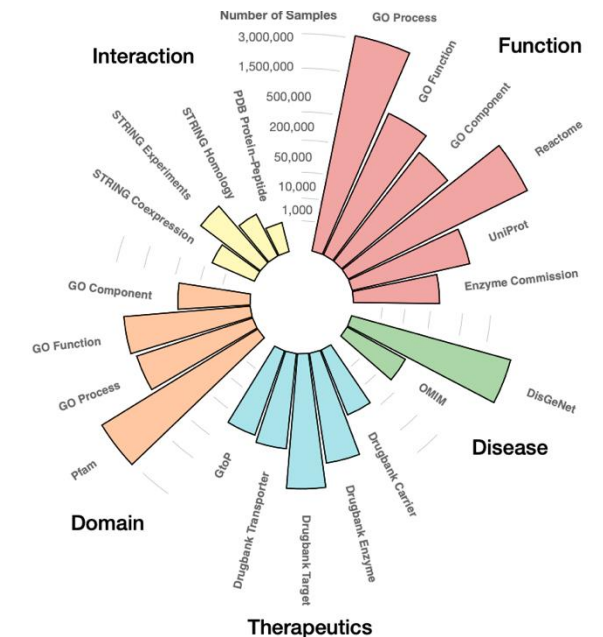
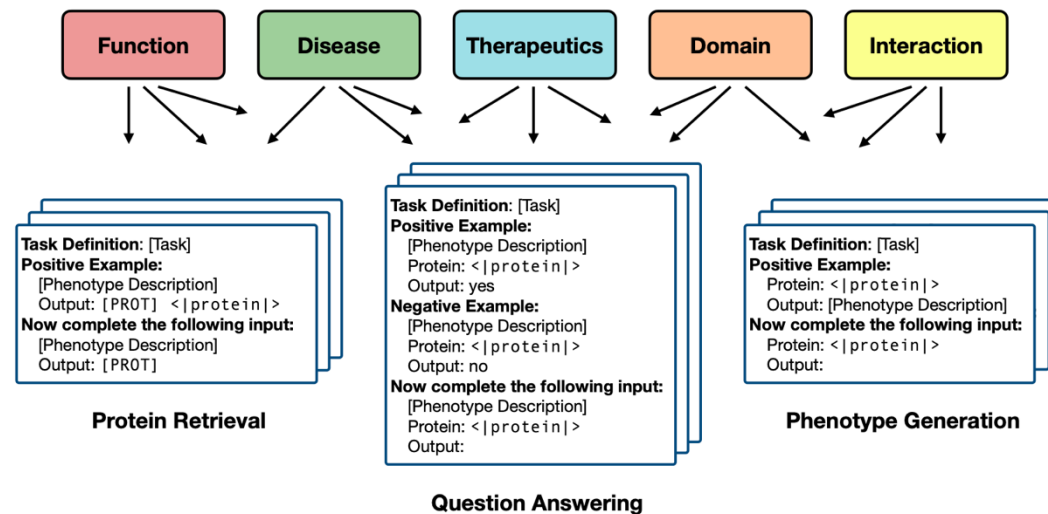


ProCyon: Foundation model for protein function



ProCyon training dataset

- 33 million protein-phenotype pairs across 5 knowledge domains
- 250 million protein sequences
- 800,000 protein structures
- 50 million full-length papers for nature language understanding



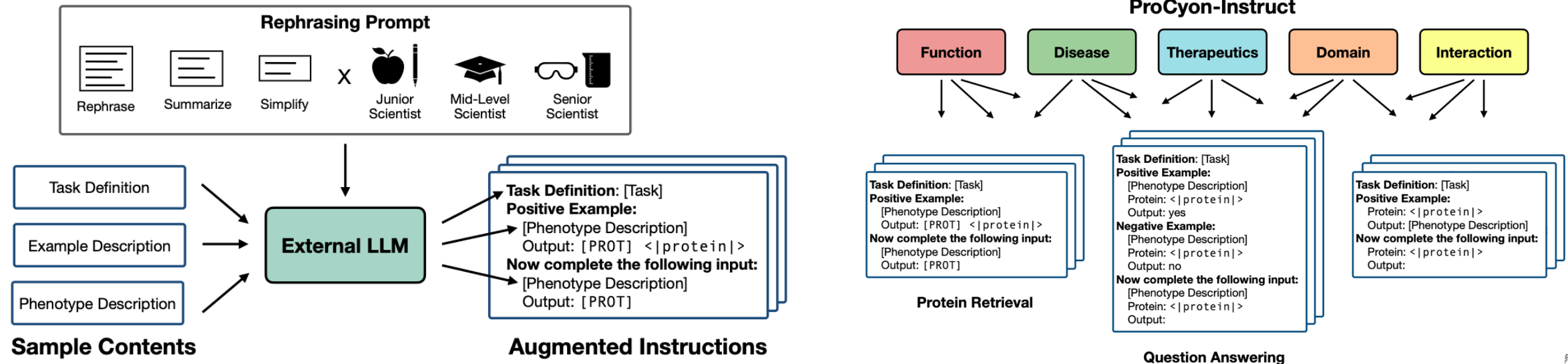
ProCyon model training



Kempner
INSTITUTE
at Harvard University

UNLOCKING
INTELLIGENCE

- Model training via three main objectives:
 - Language modeling – Next token prediction
 - Contextual protein retrieval – Contrastive learning
 - User guidance – Instruction tuning
- Final model: 11B parameter ProCyon model



Key capabilities of ProCyon

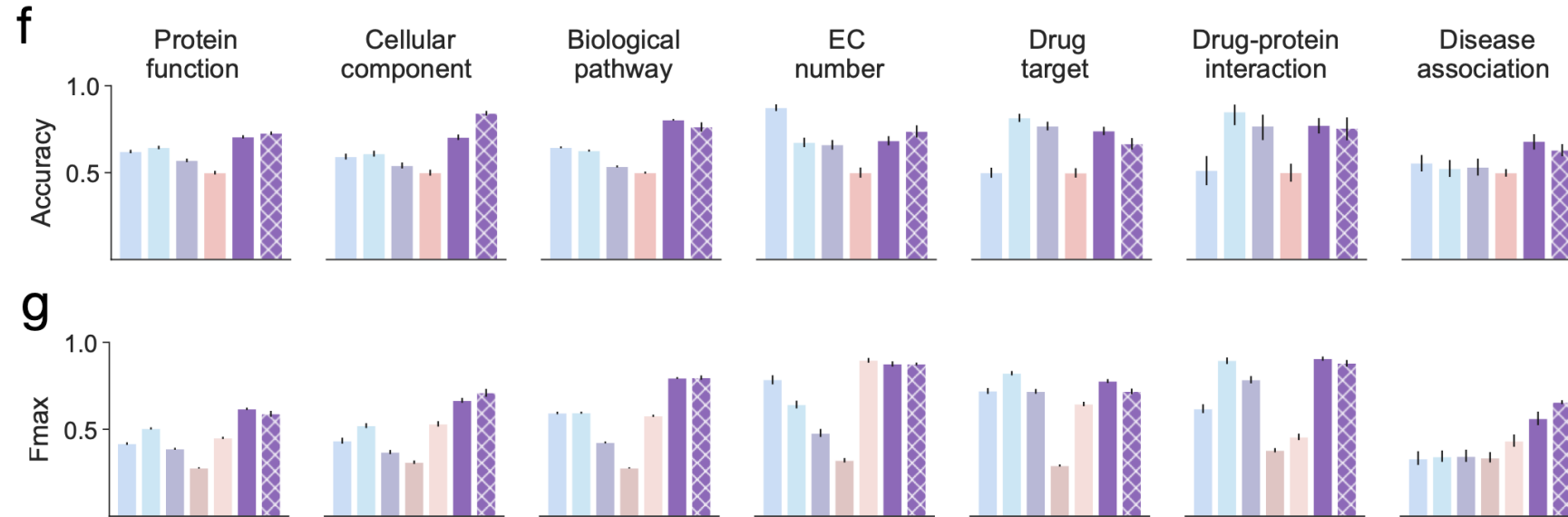
Multimodality

Zero-shot task transfer

Unbound vocabulary of user input

Prediction & generation

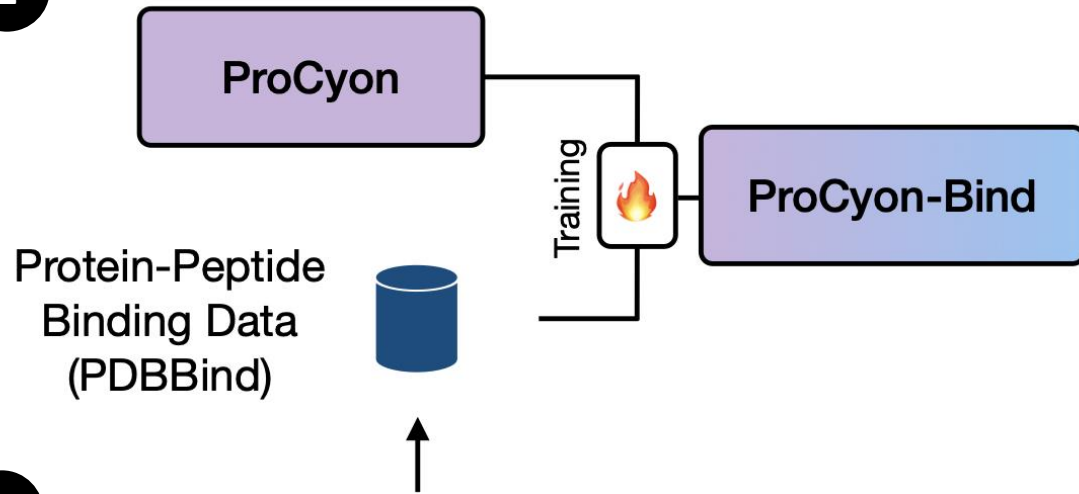
Zero-shot prediction of protein function



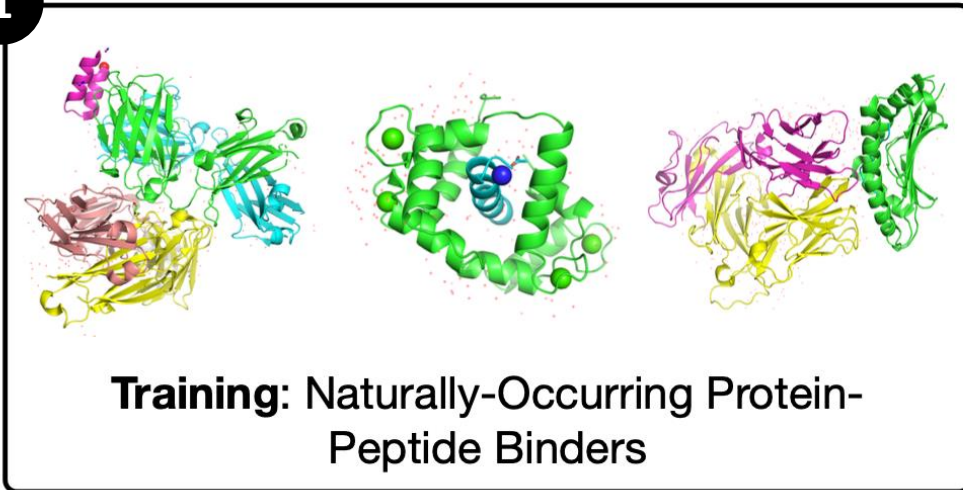
Model	Protein Sequence	Protein Structure	Free Text	Zero-shot Capable	Dynamic Task Specification
BlastKnn	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
ESM3MLP	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
GearNetMLP	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
BioTranslator	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>
ProtST	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>
ProtLLM	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>
ProCyon	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>
ProCyon (zero shot)	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>

ProCyon identifies new peptides that bind ACE2

2

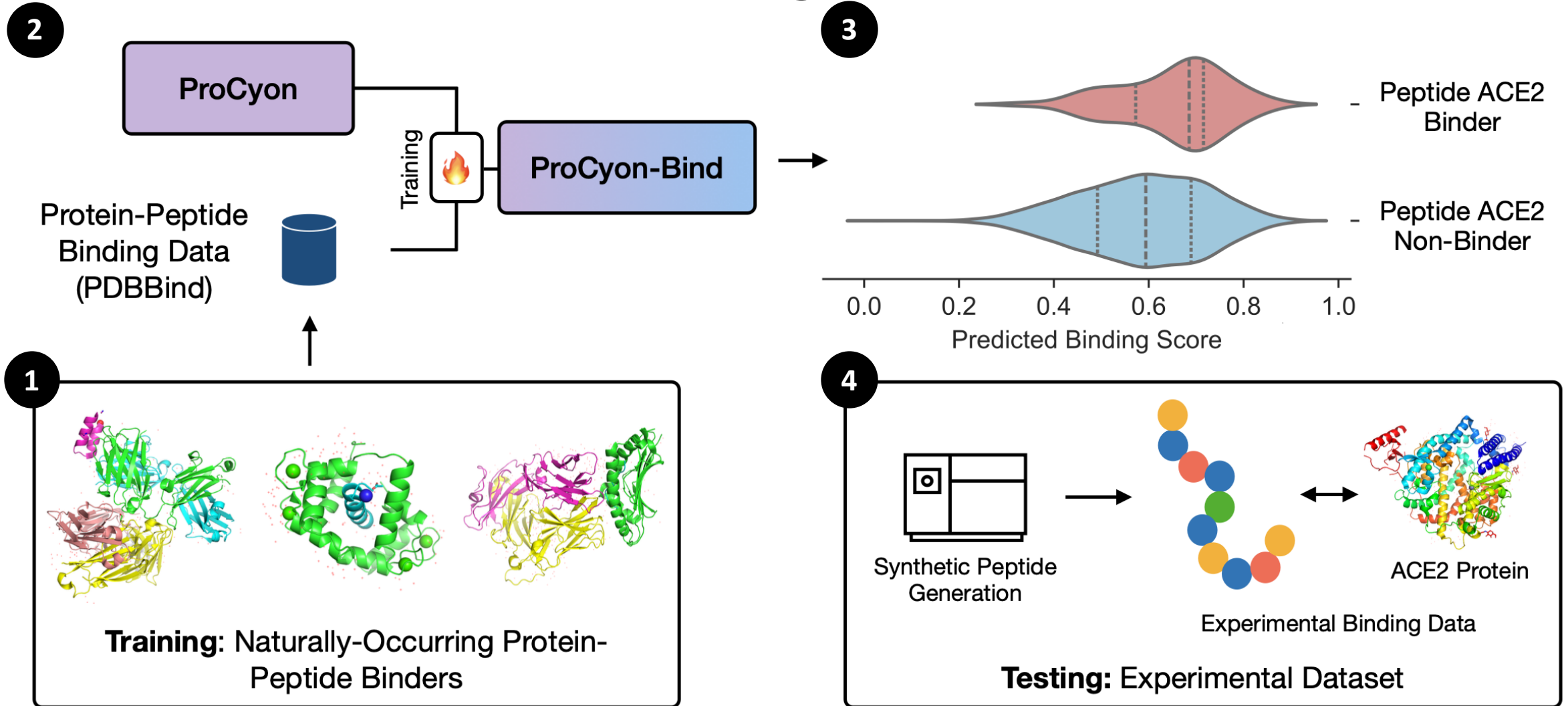


1

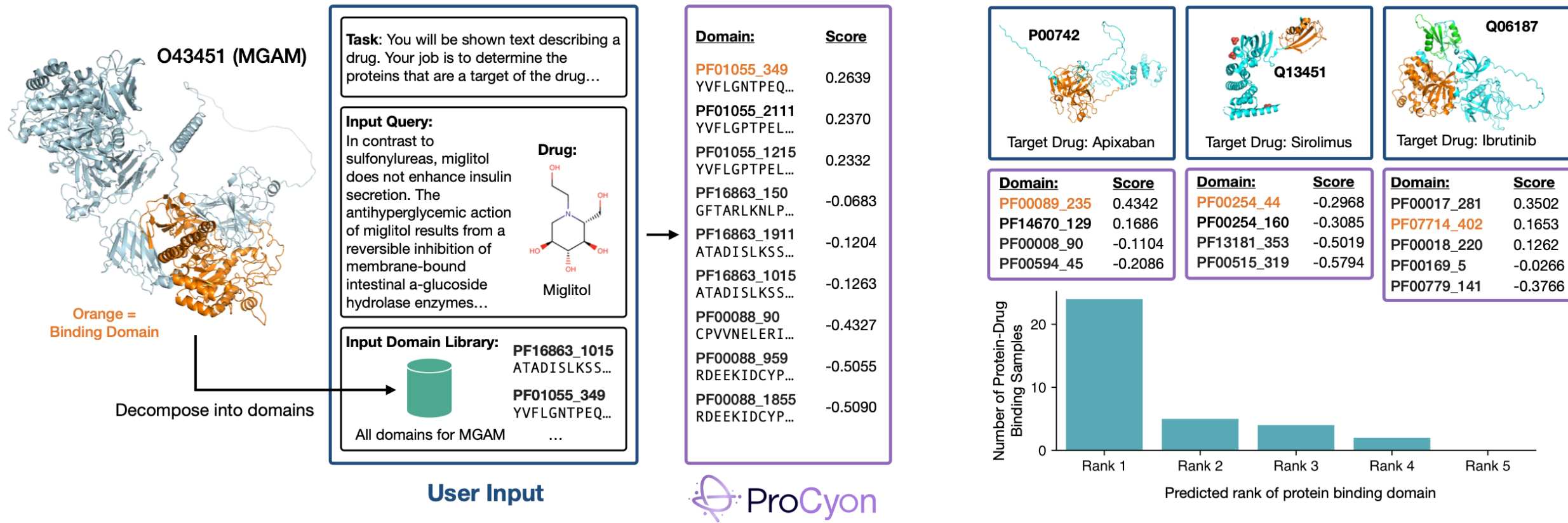


ProCyon identifies new peptides that bind

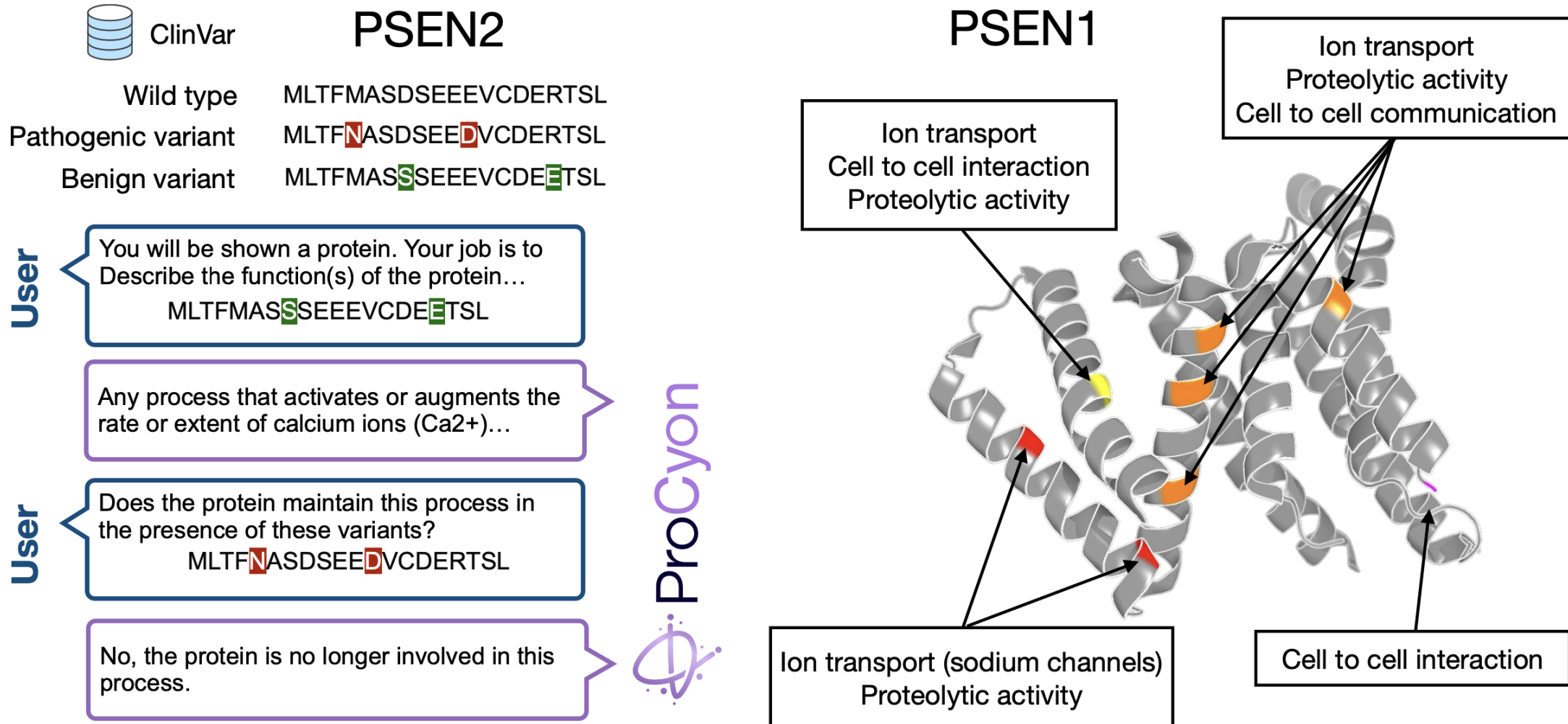
ACE2



ProCyon identifies drug-binding protein domains



Modeling phenotypes linked to genetic variants



Retrieving proteins against small molecule drugs

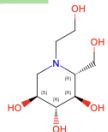
Smoking Cessation

You will be shown two texts:
one describing a disease and
one describing a drug...

Disease: Nicotine is an amine found in tobacco and tobacco products. It is the addictive agent which confers a much lower risk than other elements of tobacco, but it is not completely benign. When tobacco smoke is inhaled, nicotine rapidly enters the bloodstream through the pulmonary circulation...

Drug: Bupropion

Structure:



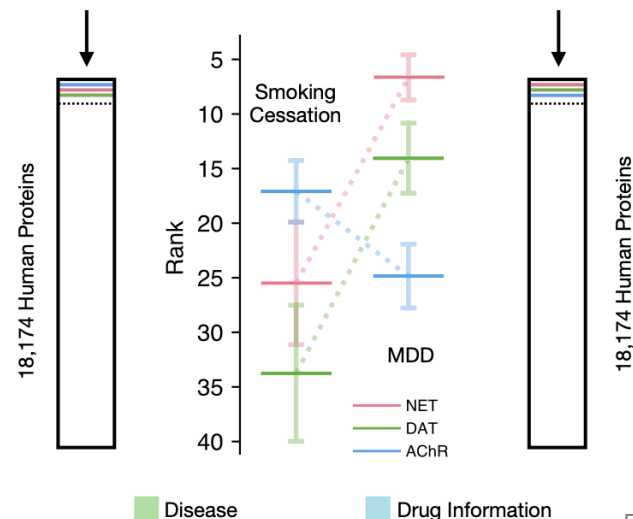
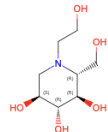
Major Depressive Disorder

You will be shown two texts:
one describing a disease and
one describing a drug...

Disease: The underlying pathophysiology of major depressive disorder has not been clearly defined. Current evidence points to a complex interaction between neurotransmitter availability and receptor regulation and sensitivity underlying the affective symptoms...

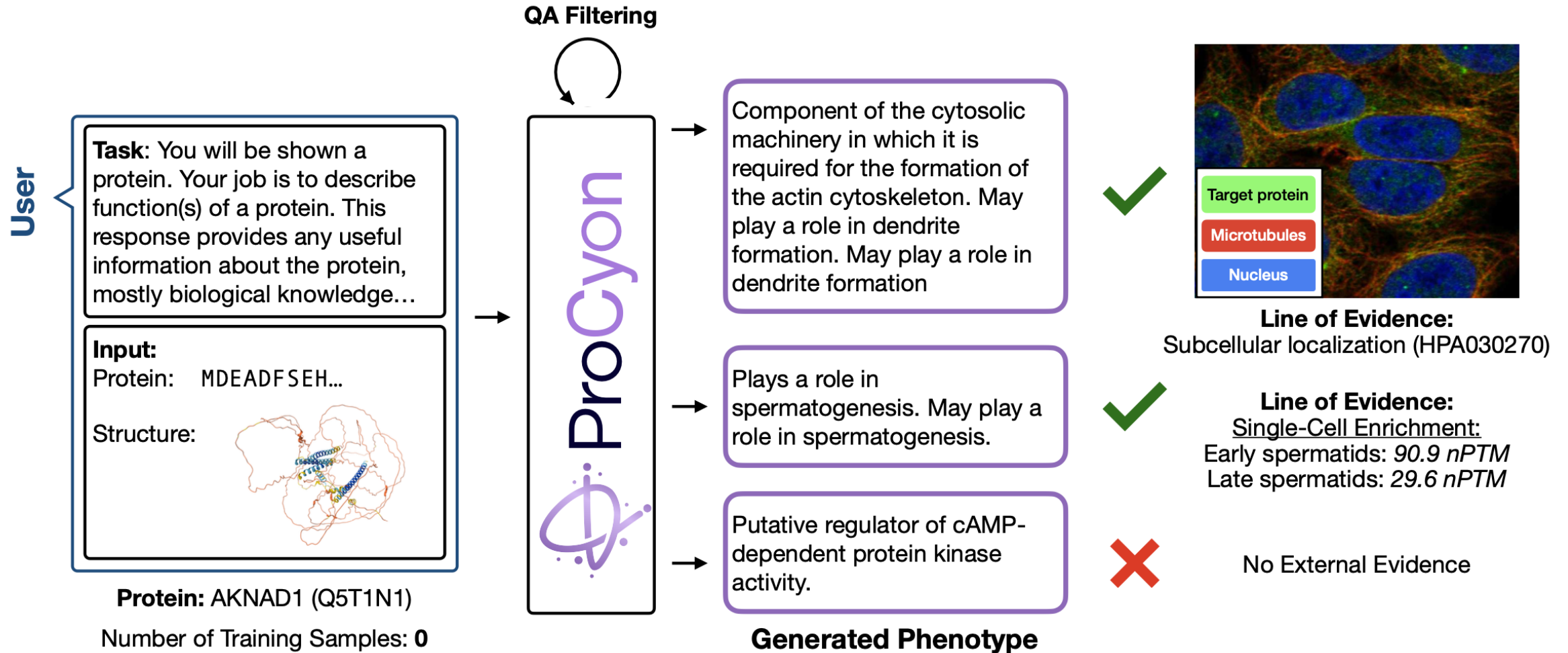
Drug: Bupropion

Structure:



- ProCyon can handle phenotype queries that interleave drug and disease knowledge domains
- One practical application of this capability is the identification of disease-specific targets, where **ProCyon can retrieve proteins targeted by a small-molecule drug for a given disease.**
- Bupropion is a small molecule drug used to treat major depressive disorder (MDD) and aid in smoking cessation
- Its **therapeutic effects are mediated through distinct protein targets:** the norepinephrine transporter (NET, UniProt: P23975), which is primarily involved in treating MDD and the cholinergic receptor nicotinic alpha 3 subunit (AChR, CHRNA3, UniProt: P32297), which plays a role in smoking cessation
- Bupropion also targets the dopamine transporter (DAT, UniProt: Q01959), which contributes to the therapeutic mechanisms of both MDD and smoking cessation

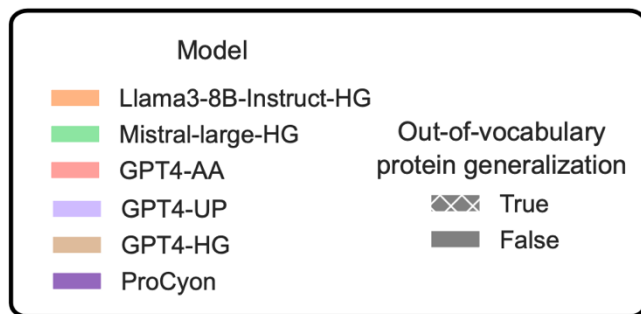
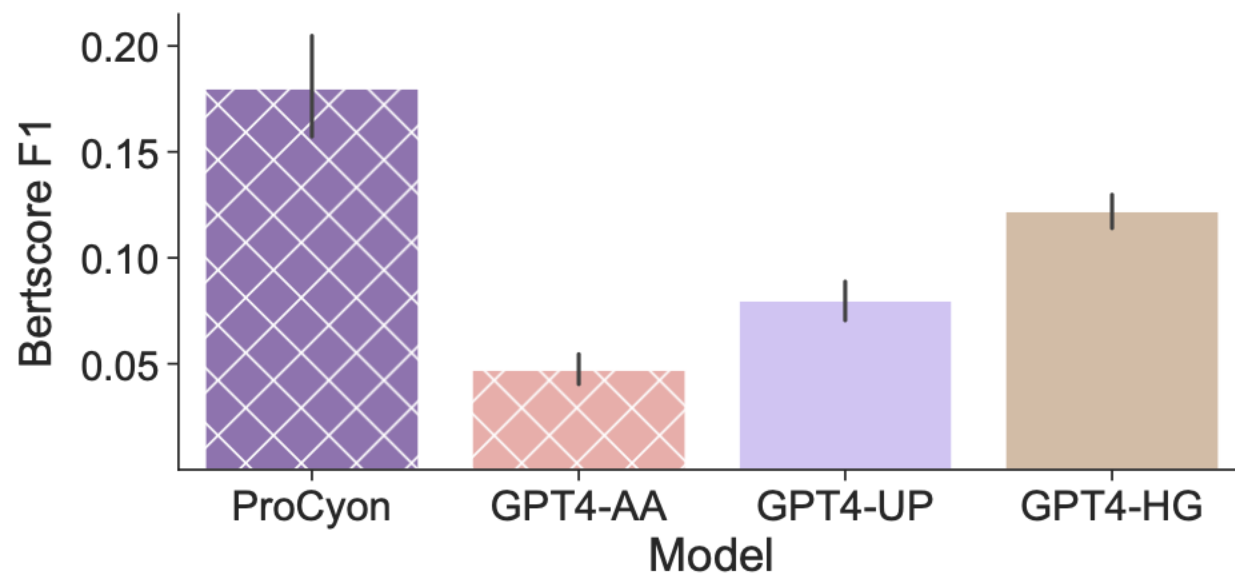
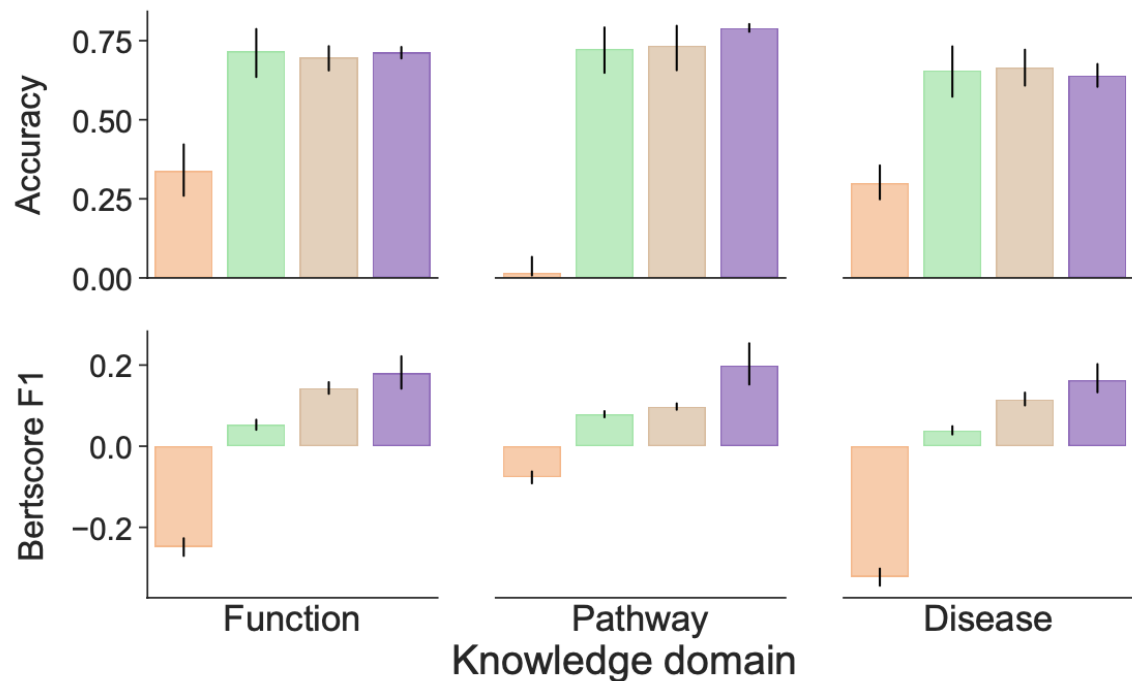
Generation of “protein caption cards”: AKNAD1



These generations are not made solely based on sequence similarity, where the closest sequence in SwissProt has only a 30% similarity

How good are generated phenotypes?

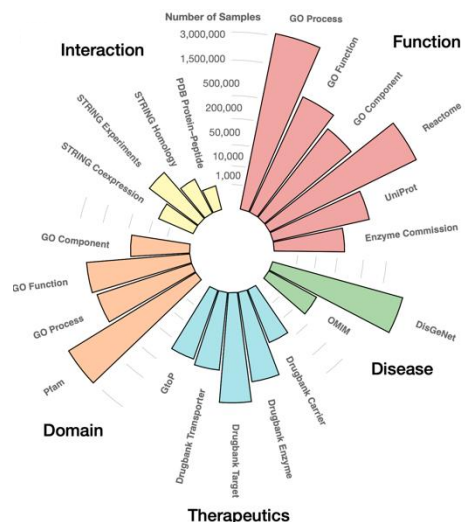
LLM-as-a-judge approach



Agentic AI: From molecules to therapies

Molecules

Knowledge-grounded AI agent for expert-level biological reasoning



Cells

Multi-agent AI for single-cell powered drug discovery



Find 5 gene candidates in goblet cells involved in non small cell lung cancer

The user are requested the top 5 genes in goblet cell involved in non-small cell lung carcinoma, `EFO: 0003060`.

The following genes in goblet cells are involved in non-small cell lung cancer:

ERBB, ALK, KRAS, BRAF, MET

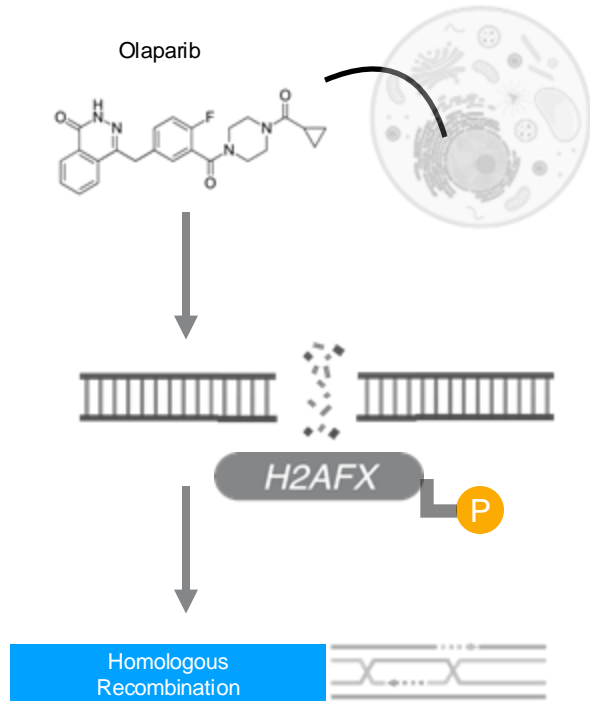
References and Code supported the analysis can be found in:

Patients

AI agent with a massive number of tools for precision therapy



AI agent for drug target identification



Prediction across hundreds of cell type and cell state contexts: Study effects of drugs across diverse cell types and cell states



Biologist

Hey Bob! Do you know any recourse or models that could help me with a step in target identification?

Sure thing! Have you try to leverage some scFoundation models #\$\$@# , additionally you can use those output and feed to some PPI networks #\$\$@# ...



Computer Scientist in Life Science



Biologist

Waite ... What? 🐹

Democratizing access to virtual cell models: Enable researchers to use in silico simulators of cell behavior and foster human-AI collaboration



Pengwei Sui



Michelle Li

AI agent for drug target identification

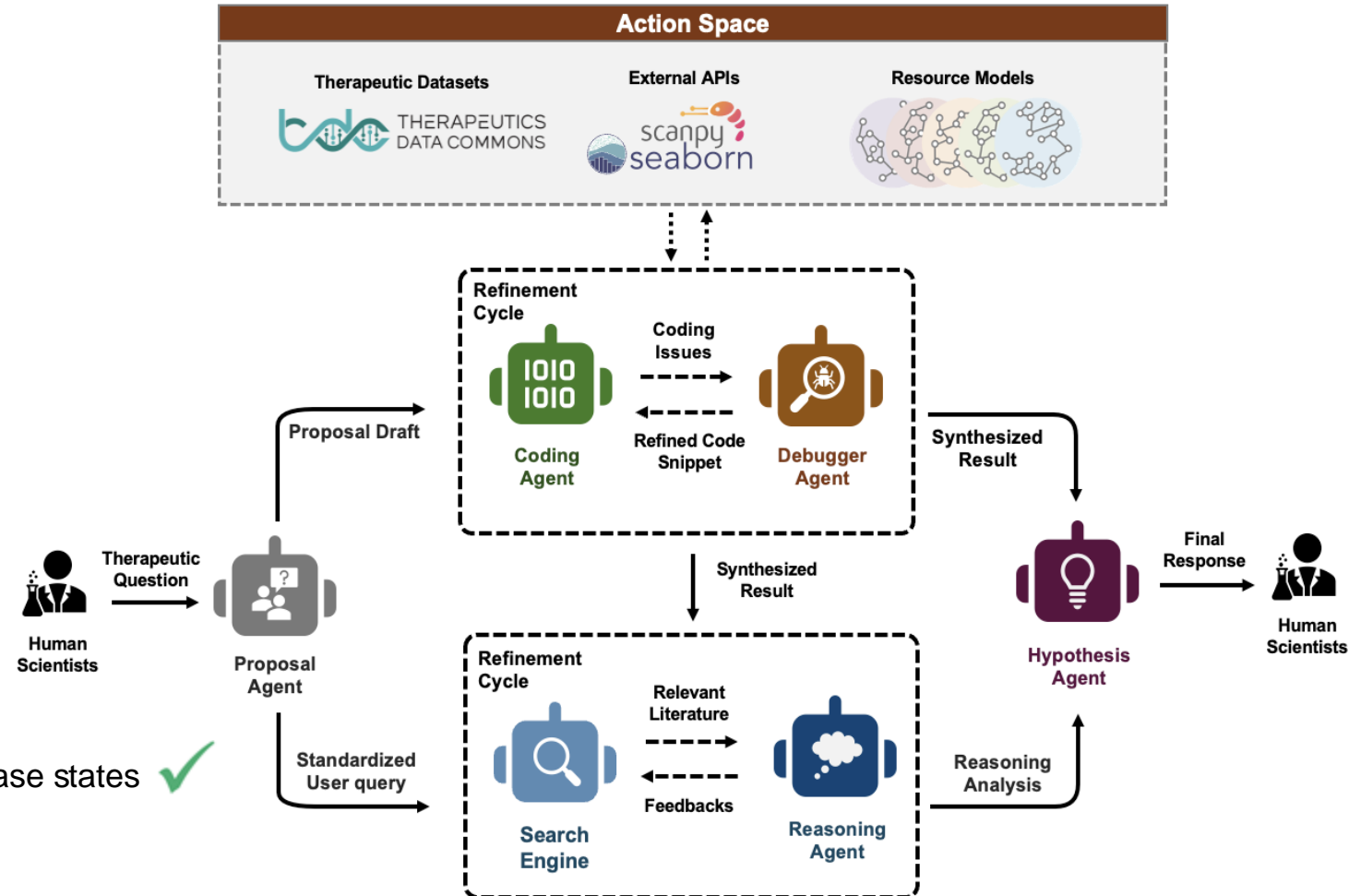
Conversational interface to single-cell atlases

Single-cell resource model knowledge

- Single-cell Genomics ✓
- Code Generation ✓
- Code Introspection & Repair ✓
- Flexible Metric Adoption ✓
- Streamline Therapeutic Data Analysis ✓
- Reasoning Through Online Literature Search ✓

Auto prediction and analysis

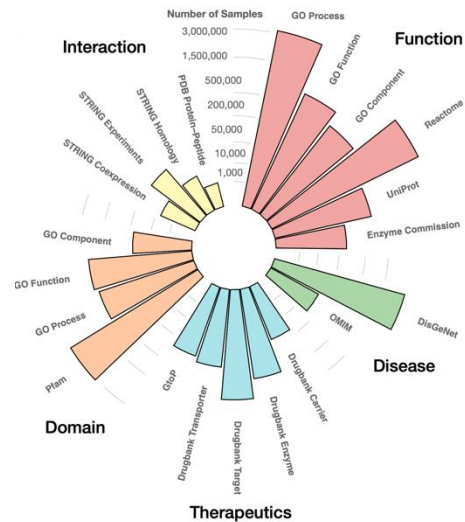
- Cell type Annotation ✓
- Gene Association Prediction ✓
- Response to chemical/genetic perturbations across disease states ✓



Agentic AI: From molecules to therapies

Molecules

Knowledge-grounded AI agent for expert-level biological reasoning



Cells

Multi-agent AI for single-cell powered drug discovery



Find 5 gene candidates in goblet cells involved in non small cell lung cancer

The user are requested the top 5 genes in goblet cell involved in non-small cell lung carcinoma, **EFO: 0003060**.

The following genes in goblet cells are involved in non-small cell lung cancer:

ERBB, ALK, KRAS, BRAF, MET

References and Code supported the analysis can be found in:

Patients

AI agent with a massive number of tools for precision therapy



Precision medicine (treatments)

Measure phenotype
and mechanisms



Design therapeutic agents
or select optimal perturbations



Provide each patient
with the right
drug, at the right
dose, at the right time

Clinical phenotypes and diseases

17,000 Diseases
7,000 Rare diseases
5-7% Rare diseases with treatments
No Treatment options for many disease subtypes

Medicines and drugs

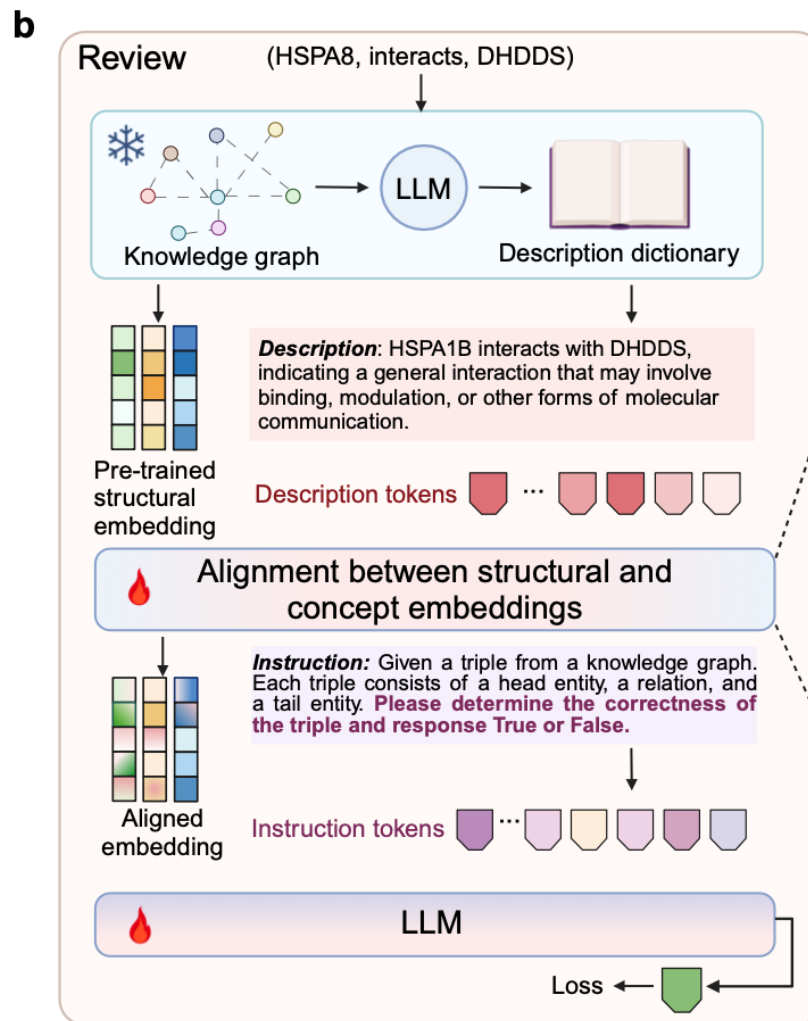
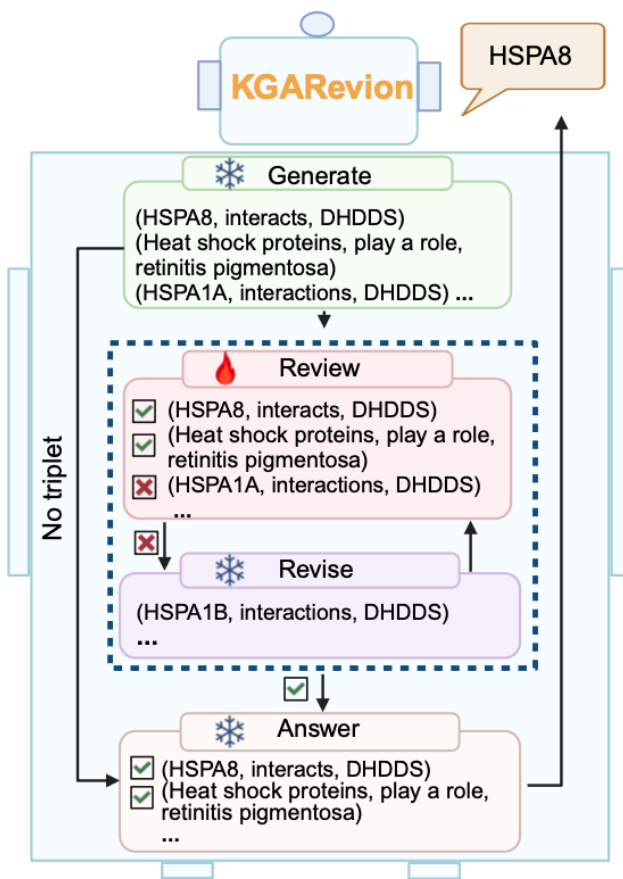
40-50 New molecules per year
30% Drugs are issued at least one post-approval new indication
Many Drugs have accrued over 10 drug indications over the years

KGARevion: KG-based LLM agent for knowledge-intensive medical QA

a

Is there an interaction between the Heat Shock Protein 70 family that acts as a molecular chaperone and the gene or protein implicated in Retinitis Pigmentosa 59 due to DHDDS mutation?

A: HSPA4 B: HSPA8 C: HSPA1B D: HSPA1A



Agent's actions:

- **Generate** candidate answers
- **Review** candidate answers:
 - Contextualized KGs
 - RAG techniques
 - Other knowledge-grounding systems
- **Revise** the answers, repeat

Objectives for this model design:

- Improved accuracy of response
- Better contextualization
- Mitigate safety and hallucination issues



Xiaorui Su

KG-based LLM agents enhance robustness and flexibility of open-ended medical reasoning

Question: A 29-year-old woman presents to the clinic with a 6-month history of progressive weakness and muscle pain. She has experienced difficulty walking and has had several falls in the past month. Her symptoms have progressed despite taking ibuprofen and acetaminophen. Physical examination reveals muscle atrophy in her upper and lower extremities. Laboratory tests show elevated creatine kinase levels and a positive test for Human Immunodeficiency Virus (HIV). What is the most likely diagnosis?

Multi-choice question

Answer Options:

- Option A: Myopathy
- Option B: Polymyositis
- Option C: Dermatomyositis
- Option D: Neuromuscular junction disorder

Changed order of answer options

Answer Options:

- Option C: Dermatomyositis
- Option A: Myopathy
- Option D: Neuromuscular junction disorder
- Option B: Polymyositis

Relabeled indices

Answer Options:

- Option G: Myopathy
- Option H: Polymyositis
- Option I: Dermatomyositis
- Option J: Neuromuscular junction disorder

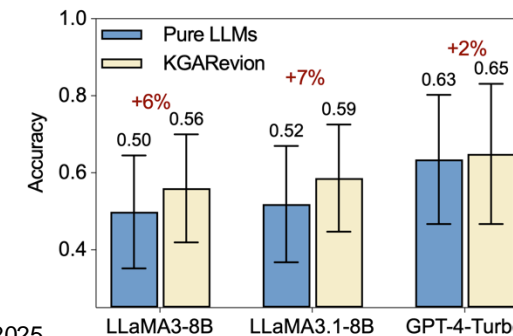
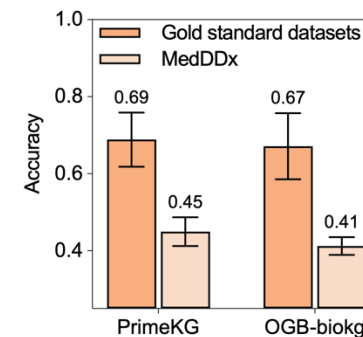
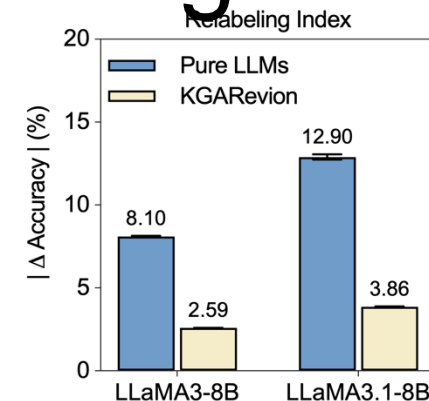
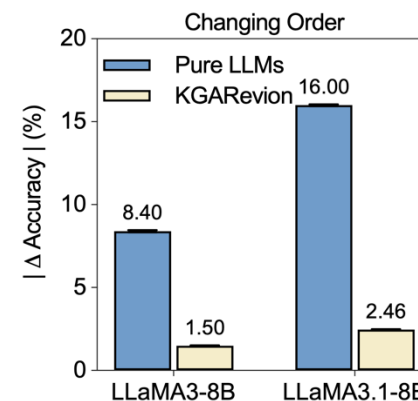
Open-ended reasoning (no options given)

Answer Options:
[EMPTY]

Robustness: KG+LLM agents are reliable and robust multiple-choice selectors

Knowledge flexibility: These systems support integration with contextualized KGs

LLM flexibility: These systems support integration with LLMs of varying sizes



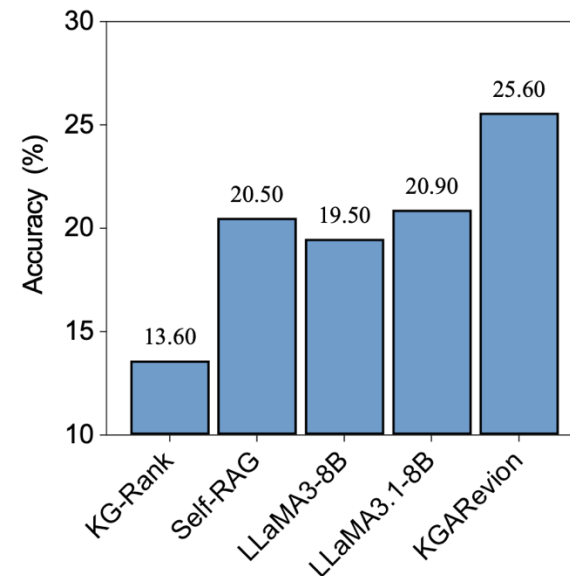
KG-based LLM agents outperform KG-only and LLM-only models on benchmarks

Multi-Choice type	Open-ended type
Which of the following best describes the structure that collects urine in the body?	What best describes the structure that collects urine in the body?
A: Bladder B: Kidney C: Ureter D: Urethra	
A microbiologist is studying the emergence of a virulent strain of the virus. After a detailed study of the virus and its life cycle, he proposes a theory: Initially, a host cell is co-infected with 2 viruses from the same virus family. Which of the following viruses is capable of undergoing the above-mentioned process?	A microbiologist is studying the emergence of a virulent strain of the virus. After a detailed study of the virus and its life cycle, he proposes a theory: Initially, a host cell is co-infected with 2 viruses from the same virus family. Which virus is capable of undergoing the above-mentioned process?
A: Epstein-Barr virus B: Human immunodeficiency virus C: Rotavirus D: Vaccinia virus	

Open-ended evaluation on MMLU-Med, MedQA-US, PubMedQA, BioASQ-Y/N, and three DDx benchmarks designed at Basic, Intermediate, and Expert difficulty levels

Method	Medical QA Benchmarks				Differential diagnosis Benchmark (DDx)		
	MMLU-Med	MedQA-US	PubMedQA*	BioASQ-Y/N	Basic	Intermediate	Expert
LLaMA3.1-8B	0.677	0.563	0.596	0.687	0.434	0.368	0.306
LLaMA3.1-8B (CoT)	0.681	0.549	0.600	0.706	0.439	0.393	0.322
LLaMA2-13B	0.442	0.253	0.252	0.455	0.286	0.338	0.317
LLaMA2-13B (CoT)	0.415	0.354	0.232	0.422	0.309	0.263	0.243
QAGNN	0.317	0.450	0.439	0.644	0.295	0.265	0.253
JointLK	0.288	0.472	0.468	0.640	0.247	0.250	0.244
Dragon	0.319	0.475	0.472	0.646	0.286	0.247	0.240
Self-RAG (7B)	0.322	0.380	0.534	0.594	0.238	0.199	0.224
Self-RAG (13B)	0.502	0.408	0.331	0.646	0.249	0.290	0.266
KG-Rank (13B)	0.452	0.362	0.305	0.503	0.253	0.256	0.234
KG-RAG (8B)	0.516	0.343	0.429	0.662	0.434	0.413	0.391
MedRAG (70B)	0.579	0.487	0.574	0.719	0.365	0.348	0.327
KGAREVION (LLaMA3, $k = 1$)	0.703	0.610	0.562	0.744	0.473	0.404	0.395
KGAREVION (LLaMA3, $k = 3$)	0.678	0.628	0.590	0.737	0.469	0.451	0.411
Improvement over best baseline	+5.2%	+6.2%	+0.4%	+6.3%	+3.9%	+8.3%	+3.2%
KGAREVION (LLaMA3.1, $k = 1$)	0.734	0.618	0.619	0.763	0.483	0.457	0.409
KGAREVION (LLaMA3.1, $k = 3$)	0.716	0.620	0.638	0.749	0.469	0.411	0.447
Improvement over best baseline	+5.3%	+5.7%	+3.8%	+4.4%	+4.4%	+4.4%	+5.6%

AFRIMED-QA

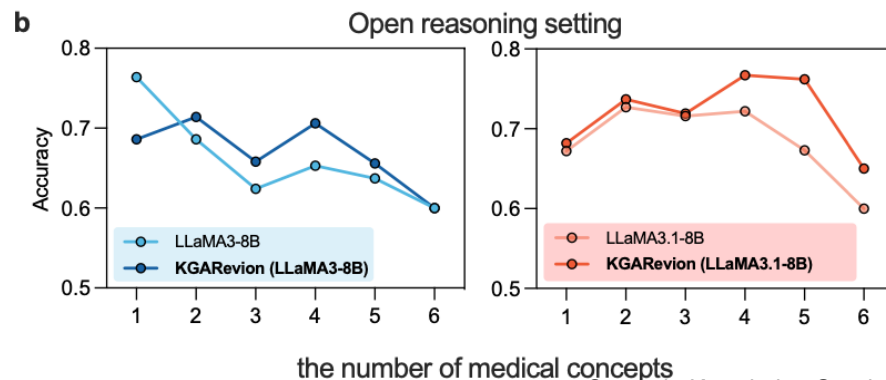
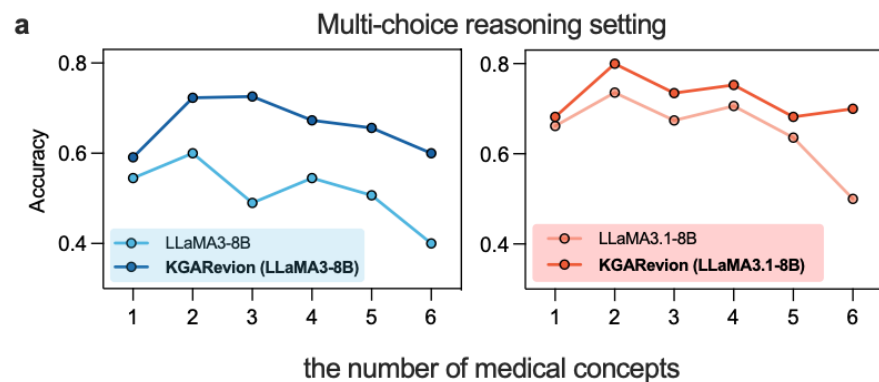


Benchmarking completed after the knowledge cut-off data for LLM and the KG

Reason about knowledge-intensive questions

Question: A 29-year-old woman presents to the clinic with a 6-month history of **progressive weakness** and **muscle pain**. She has experienced **difficulty walking** and has had **several falls** in the past month. Her symptoms have progressed despite taking **ibuprofen** and **acetaminophen**. Physical examination reveals **muscle atrophy** in her upper and lower extremities. Laboratory tests show elevated **creatinine kinase** levels and a positive test for **Human Immunodeficiency Virus** (HIV). What is the most likely diagnosis?

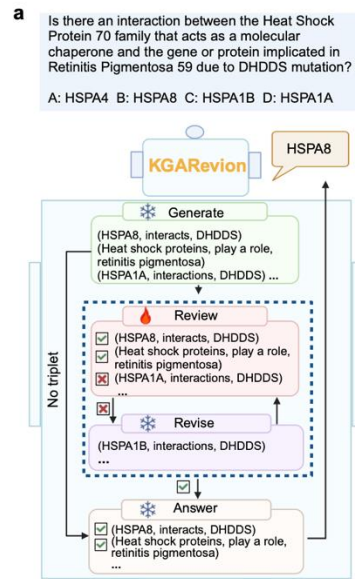
Clinical concepts



Reason about knowledge-intensive questions

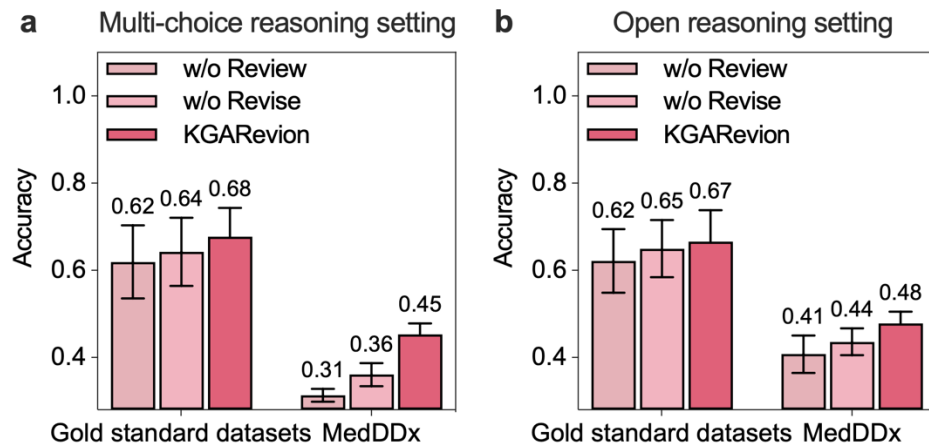
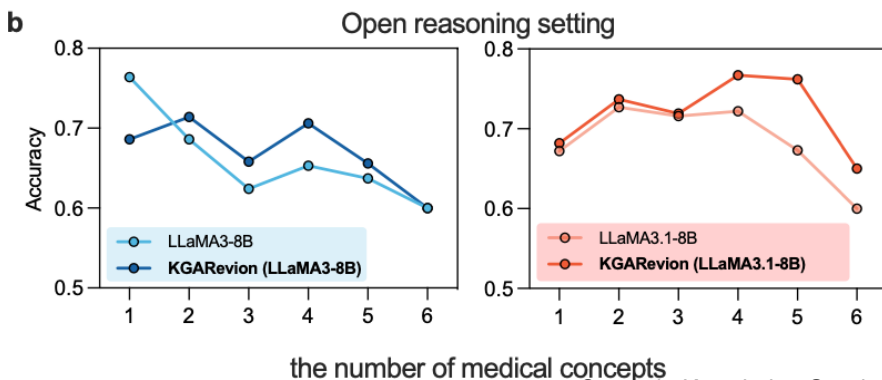
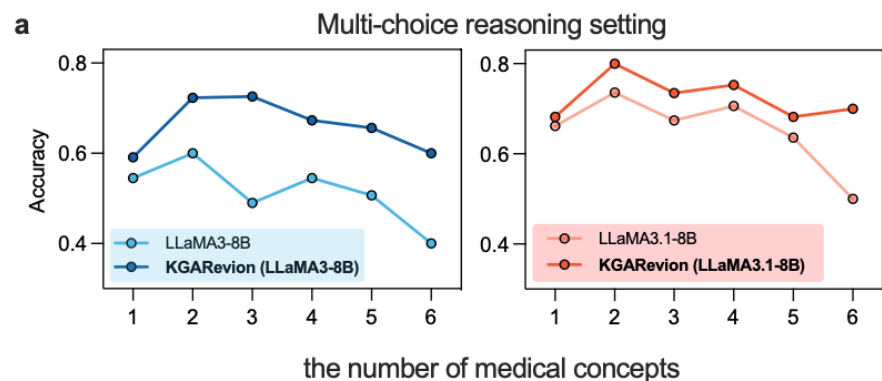
Question: A 29-year-old woman presents to the clinic with a 6-month history of **progressive weakness** and **muscle pain**. She has experienced **difficulty walking** and has had **several falls** in the past month. Her symptoms have progressed despite taking **ibuprofen** and **acetaminophen**. Physical examination reveals **muscle atrophy** in her upper and lower extremities. Laboratory tests show elevated **creatinine kinase** levels and a positive test for **Human Immunodeficiency Virus (HIV)**. What is the most likely diagnosis?

Clinical concepts

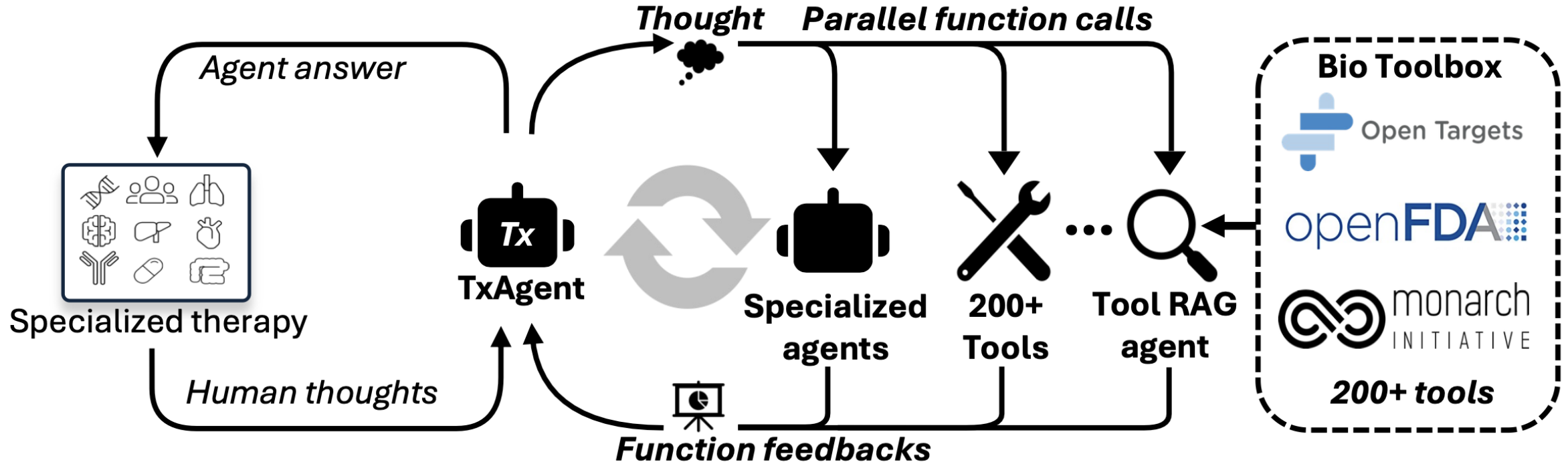


Do these actions matter? - Yes

- **Generate** candidate answers
- **Review** candidate answers:
 - Contextualized KGs
 - RAG techniques
 - Knowledge-grounding systems
- **Revise** the answers, repeat

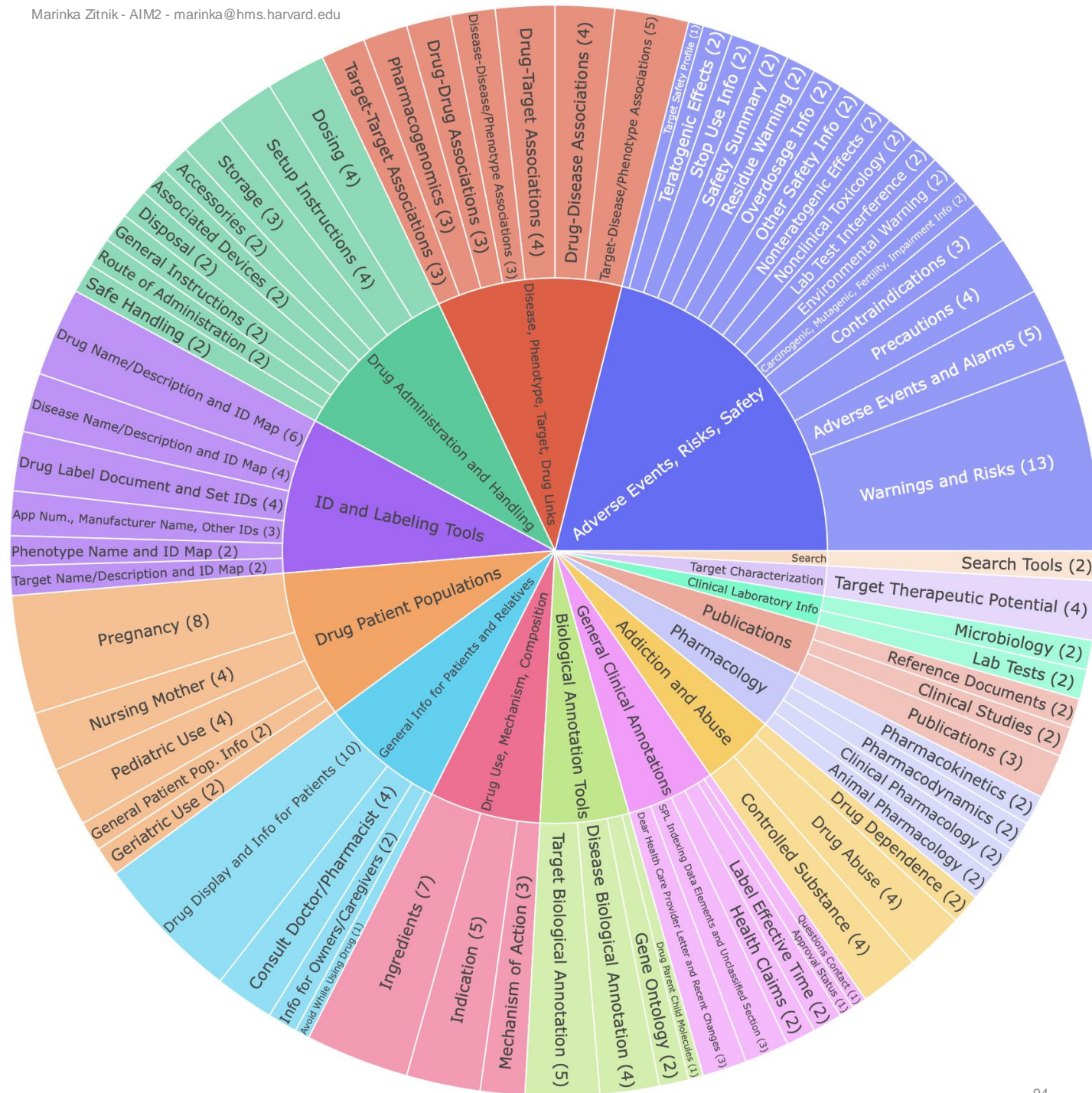


TxAgent: Precision therapy agent using massive number of tools



Shanghua Gao

Agent is trained to use 200+ tools!



TxAgent's toolbox



Open Targets

- Targets, drugs, diseases:
 - Inter-entity associations, symptoms, indications, etc.
- 56 tools



- FDA drug product labels:
 - Indications, contraindications, adverse events, geriatric effects, etc.
- 151 tools



monarch
INITIATIVE

- Phenotype, symptom, and disease associations
- 3 tools

TxAgent's toolbox



Open Targets

{**name:** get_associated_targets_by_disease_efoID
description: Find targets associated with a specific disease or phenotype based on EFO ID.
input: EFO ID}



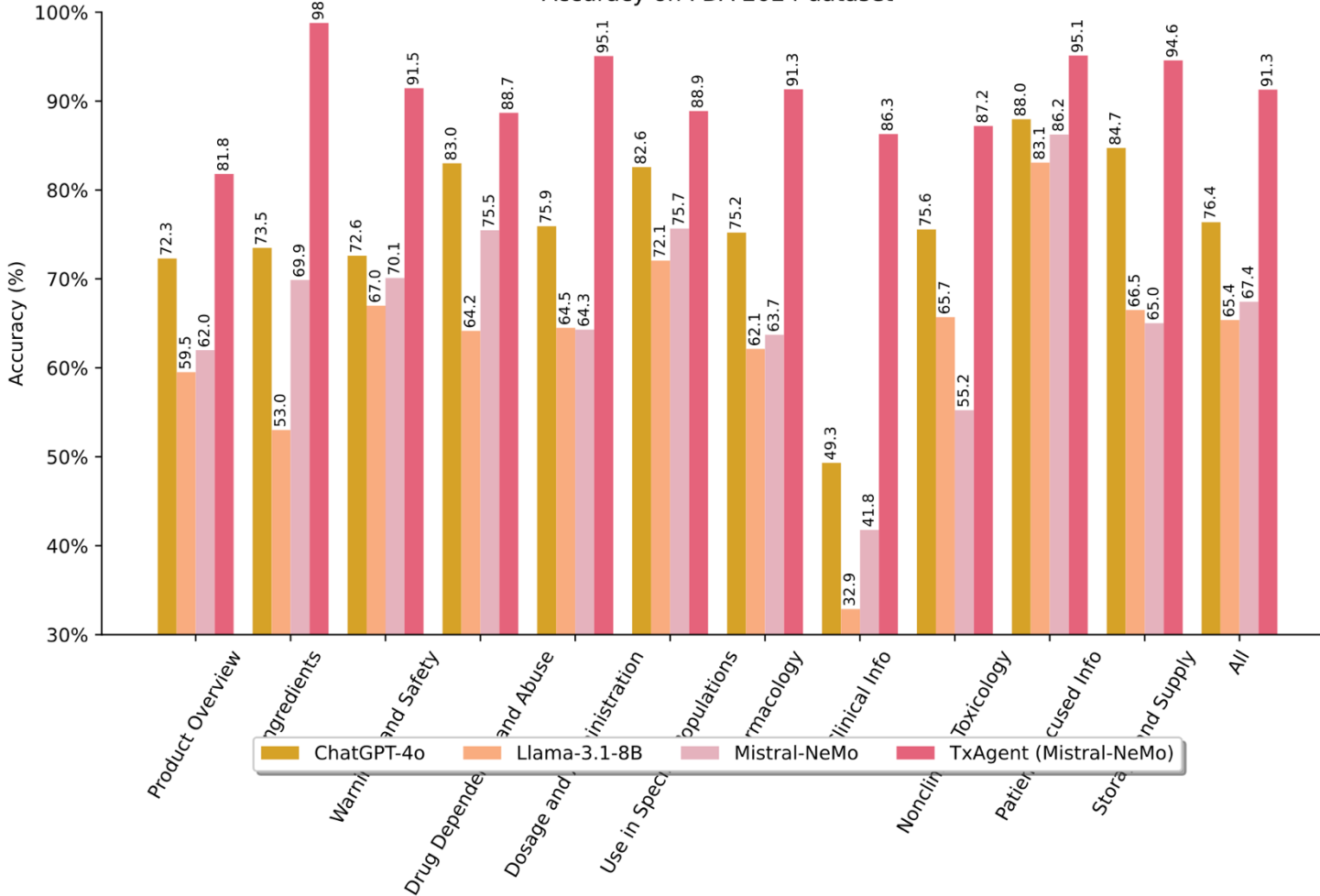
{**name:** get_contraindications_by_drug_name
description: Retrieve contraindications information based on the drug name.
input: Drug name}



{**name:** get_associated_diseases_by_HPO_ID
description: Retrieve diseases associated with a list of phenotypes or symptoms by the phenotype HPO IDs.
input: HPO ID}

TxAgent performance

Accuracy on FDA 2024 dataset



New Drug

What serious adverse event has been observed in patients treated with monoclonal antibodies like **Kisunla**?

TxAgent

.....The most common adverse reactions include **Amyloid Related Imaging Abnormalities (ARIA-E, ARIA-H microhemorrhage, ARIA-H superficial siderosis), headache, and infusion-related reactions.....three patients (0.4%) treated with Kisunla had serious adverse reactions of intestinal obstruction, and two patients (0.2%) had serious adverse reactions of intestinal perforation.**

- Latest information from FDA. ✓
- Serious adverse event with clear reason. ✓

ChatGPT-4o

A serious adverse event observed in patients treated with **monoclonal antibodies like Kisqali (ribociclib)** is severe hepatotoxicity (liver damage). Other significant risks include **neutropenia (a drop in white blood cells), QT prolongation (heart rhythm issues), and pulmonary toxicity (lung-related complications).**

.....

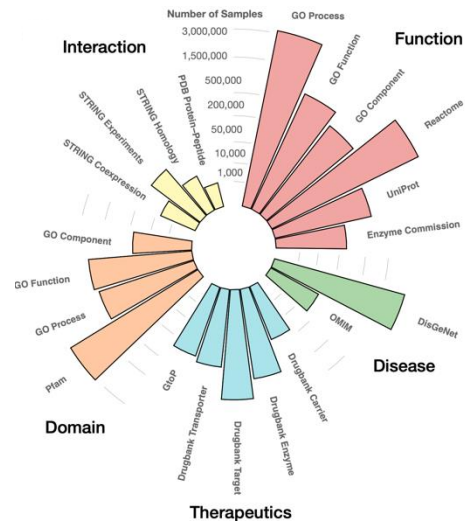
- Hallucinates to Kisqali. ✗
- Wrong adverse event without convincing explanation. ✗

New drug approved by FDA in July to treat Alzheimer's disease.

Agentic AI: From molecules to therapies

Molecules

Knowledge-grounded AI agent for expert-level biological reasoning



Cells

Multi-agent AI for single-cell powered drug discovery



Find 5 gene candidates in goblet cells involved in non small cell lung cancer

The user are requested the top 5 genes in goblet cell involved in non-small cell lung carcinoma, **EFO: 0003060**.

The following genes in goblet cells are involved in non-small cell lung cancer:

ERBB, ALK, KRAS, BRAF, MET

References and Code supported the analysis can be found in:

Patients

AI agent with a massive number of tools for precision therapy



Today's lecture

1. What are agents?
2. A brief history of LLM-powered agents in the context of LLMs and agents
3. Overview of LLM-powered agents
4. Tool use and learning
5. Examples of agents in biology and medicine