

---

# Integrating Internal Cognitive Tools and External Physical Tools for Language Agent

---



Hongru WANG

<https://rulegreen.github.io/>

The Chinese University of Hong Kong

Supervisor: Prof. Kam-Fai Wong



# Background



王泓睿

Hongru WANG

Ph.D. Candidate

Department of Systems Engineering and  
Engineering Management,  
The Chinese University of Hong Kong



- **Google Scholar:** Citations: 380+, h10-index: 13
- **Best Paper Awards:** SIGHAN @ACL 2024, International Doctoral Forum
- **Publications:** NeurIPS (1), ACL (4), EMNLP (11), NAACL (3), COLING (3), ICASSP (2), Computer Speech & Language (1), Machine Intelligence (1)
- **Representative works:** Cue-CoT, SAFARI, SelfDC, AppBench
- **First Tool Learning Tutorial @ SIGIR 2024**
- **First Cantonese-based Task-oriented Dialogue System (KddRES)**
- **Champion on WWW2024 Online Safety Prize Challenge**
- **Funding and Grants:** TBF22ENG004 and OSCP 2023-2024
- **Interns and Visiting:** UIUC, EdinburghNLP, ByteDance
- **Co-founder and Committee Member:** NLP Academic Exchange Platform (NICE)
  - Homepage: <https://nice-nlp.github.io/>
  - Fans: 10w+; Talks: 60+; Invited Speakers: 100+; Views: 15w+
- **Research Interests:** Dialogue System, Tool Learning, Language Agent

# Overview

- Introduction
- Methods
  - Internal Cognitive Tools
  - External Physical Tools
  - Frameworks (Tool Planning)
- Benchmarks
  - Single-turn Interaction
  - Multi-turn Interaction
- Summary and Future Directions

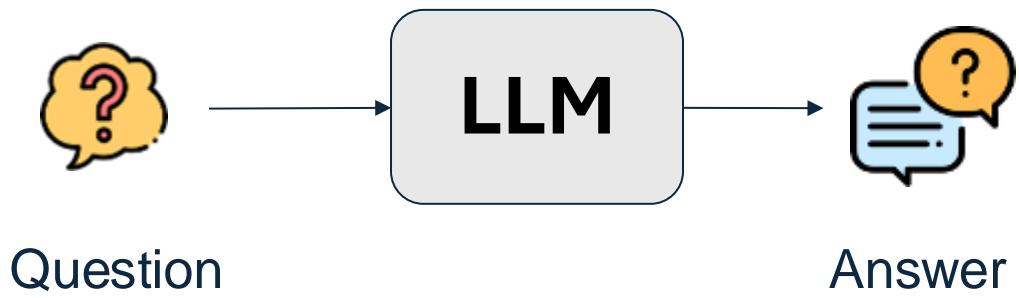


# Overview

- Introduction
- Methods
  - Internal Cognitive Tools
  - External Physical Tools
  - Frameworks (Tool Planning)
- Benchmarks
  - Single-turn Interaction
  - Multi-turn Interaction
- Summary and Future Directions



# Introduction



A typical usage of LLMs

Exceptional performance and can almost answer any types of problems

However, still

lack of **reasoning** explainability and reliability  
(personalization, trustworthy, safety, ...)

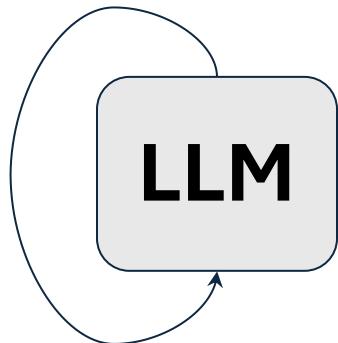
lack of **acting** with external environment (up-to-date knowledge, domain-specific service, ...)

lack of **planning** to decide when to reason and when to act, and in what ways (decision-making in complex environment)

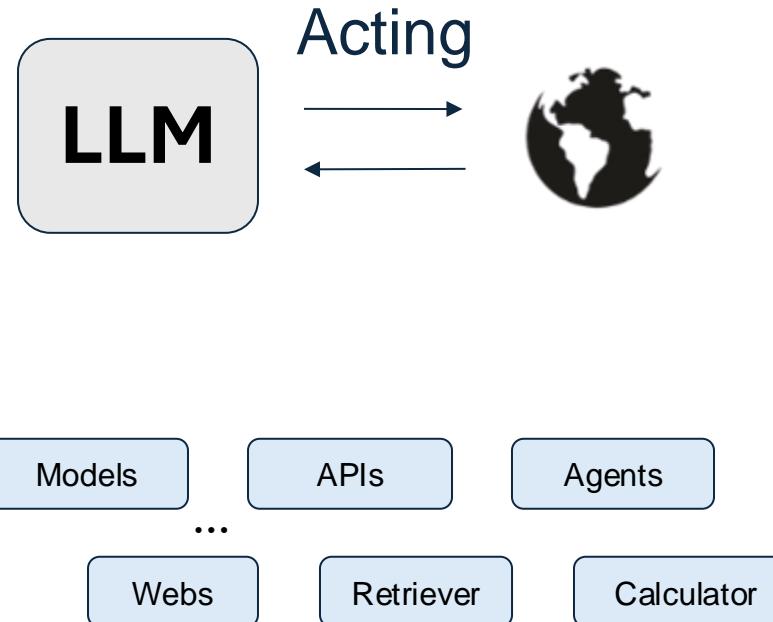


# Reasoning vs Acting vs Planning

## Reasoning



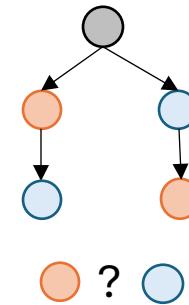
CoT      ...      Decomposition  
...  
Backward      Reflection      Alternative



Models      ...      APIs      Agents  
...  
Webs      Retriever      Calculator

## Planning

Reason or Act



Cognitive mechanism / functions  
from Cognitive Science

Physical Tools / Actions

Decision-making  
Planning



# Reasoning vs Acting vs Planning



Shunyu Yao  
@ShunyuYao12

...



reasoning == acting ?

If reasoning == acting [Yao et al, ...]

then how should we view them in an unified manner?



else:

**My Answer: reasoning and acting are both tools**

what's real difference between reasoning and acting?



# What's Tool Anyway?

what do you think tool is?

A tool is an implement, device, or piece of software used to carry out a specific function or task. Tools can be physical, like a hammer or a screwdriver, or digital, like a software application or online service. Their primary purpose is to make tasks easier, more efficient, or possible to accomplish. Tools can be simple or complex and can be used in a wide range of fields, including construction, computing, education, healthcare, and more.

Claude-3-Haiku Poe

"Tool" is a broad term that can refer to various objects or devices used to perform a specific task or function. In a general sense, a tool is an instrument or implement designed to help a person accomplish a particular goal or activity more efficiently or effectively.

Tools can come in many different forms, such as:

1. Handheld implements: These are tools that are held and manipulated by hand, like hammers, screwdrivers, pliers, saws, etc. They are commonly used in construction, repair, and maintenance work.
2. Powered tools: These tools use some form of power source, such as electricity, batteries, or fuel, to assist or automate a task. Examples include power drills, chainsaws, lathes, and sanders.
3. Measuring tools: These tools are used to quantify or assess various properties, such as length, weight, temperature, or time. Examples include rulers, scales, calipers, and stopwatches.
4. Computer software and applications: In the digital age, software programs and apps on computers, smartphones, and other devices can also be considered tools. They help users perform tasks like writing, analysis, communication, and entertainment.
5. Conceptual tools: Some "tools" are not physical objects but rather mental frameworks, methodologies, or approaches that help people solve problems, make decisions, or accomplish goals. Examples include problem-solving techniques, decision-making models, and management strategies.

what do you think tool is?

Gemini-1.0-Pro Poe

## What is Tool?

Tool is a rock band formed in Los Angeles, California in 1990. They are known for their experimental and progressive music, which blends elements of metal, rock, and alternative.

Here's a breakdown of what Tool is:

**Genre:** Progressive rock, metal, alternative

**Members:**

- Maynard James Keenan (vocals)
- Adam Jones (guitar)
- Danny Carey (drums)
- Justin Chancellor (bass)

unrelated answers due to noisy information  
physical or digital to complete the task  
conceptual tools in mental framework



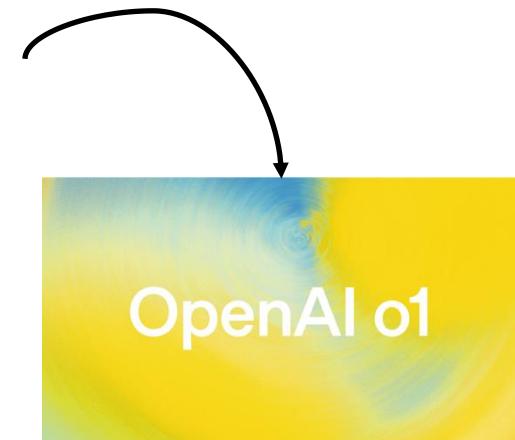
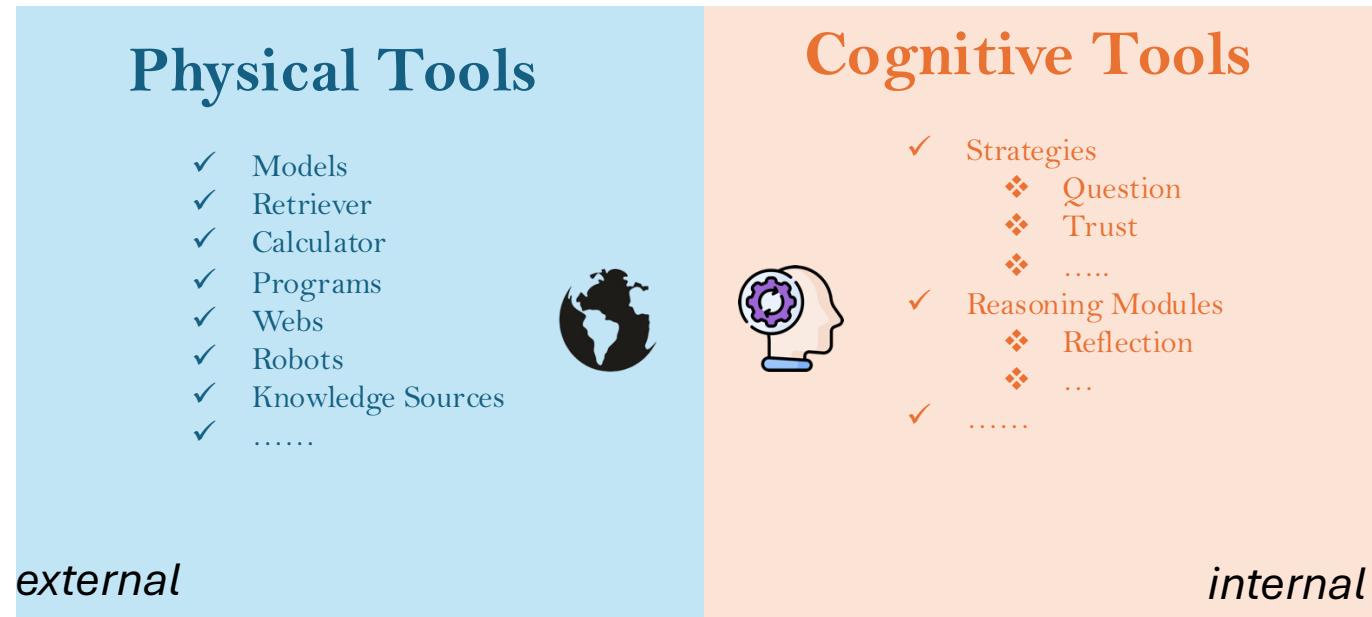
# Reasoning == Acting (in) Tools

To reason and act is the same thing

翻译帖子

下午10:56 · 24/6/24 来自 Earth · 865 次查看

**Tool** is defined as object that can extend an individual's ability to modify features of the surrounding environment or help them accomplish a particular task in general. It can be **internal cognitive/conceptual tools** and **external physical tools**.



# Reasoning == Acting (in) Tools

## Essence of Tool

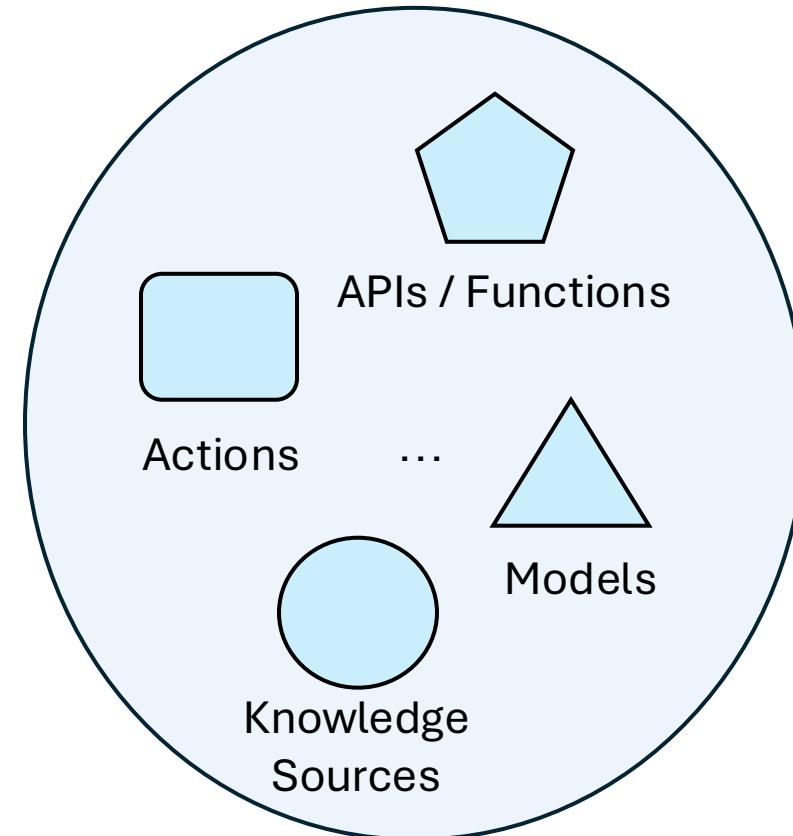
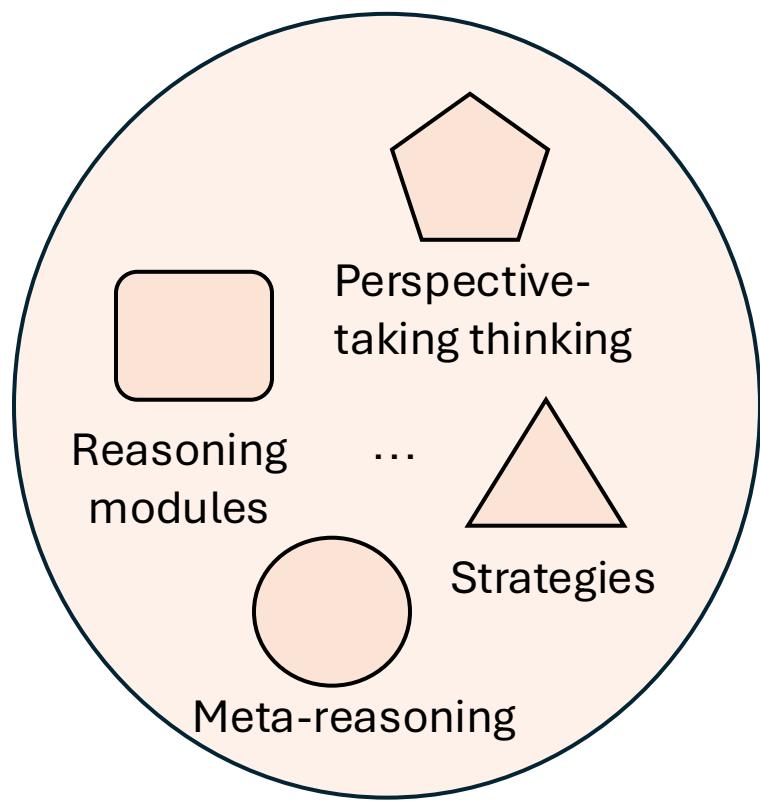
- **Useful:** A tool must effectively complete one or multiple tasks. It typically receives inputs and produces outputs.
- **On-demand:** A tool must be used as needed, meaning it is invoked based on the current task.

Use <u>tool</u> to <u>do</u>	Types	Features
calculator calculate retriever retrieve models generate / classify	physical tools	math operations up-to-date/factual multi-modal
CoT think step by step reflection reflect question ask psychology express	cognitive tools	reasoning self-critique/evolving proactivity humanity



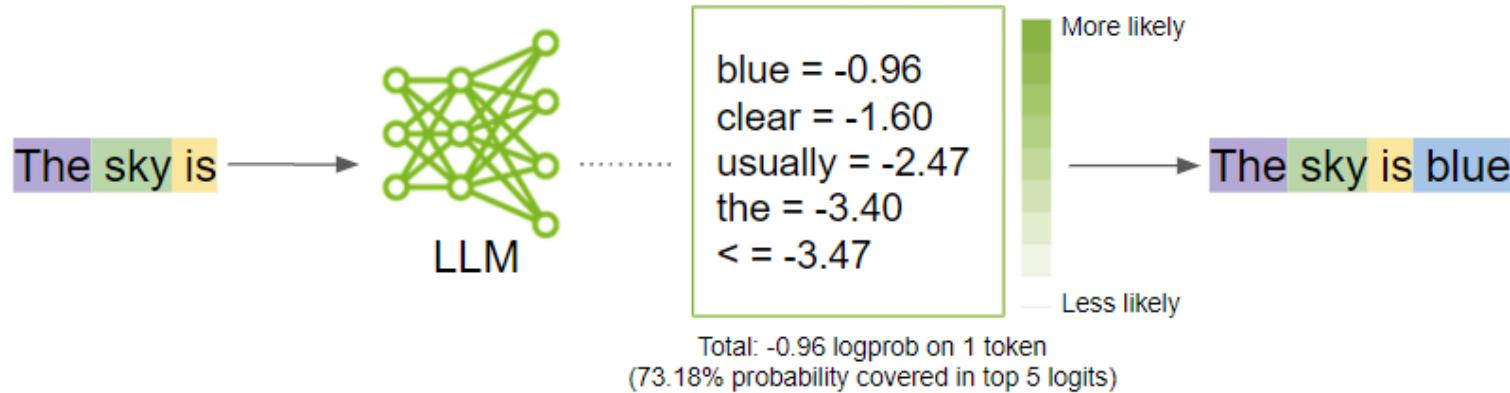
# Why This Matter ? (1)

- ❑ Reasoning is just one type of internal cognitive tools, while acting is one type of external physical tools.
- ❑ All tools either be internal cognitive tools or external physical tools. Most of interactions (if not all) can be regarded as user-agent-tools interactions.



# Why This Matter ? (2)

- **Next Token Prediction:** fundamental task to learn (or unify) almost all natural language tasks and world knowledge, including understanding and generation.



- **Next Tool Prediction:** *maybe* one of fundamental tasks to learn (or unify) almost **all interactions** and **gain new world knowledge**.

Or at least we need a unified, fundamental, practical solution to learn all of interactions either internally or externally, just as we observed in next token prediction.



# From LLMs to Language Agents

Reasoning



Planning: Reason or Act



Cognitive Tools  
Physical Tools  
Tool Planning  
Evaluation  
.....

Language Agent

- ✓ personalized and trustworthy
  - ✓ flexible and generalizable
  - ✓ practical and helpful
- .....



# Cognitive Tools

- Chain-of-Cues
- Perspective-taking Thinking
- Meta-reasoning Theory
  - Process Supervision
  - Confidence monitoring
  - .....



*internal*

# Physical Tools

- Knowledge Source as Tools (i.e., RAG)
- Tool Creation (i.e., Retriever)
- Actions as Tools (i.e., Embodied AI)



*external*



# Internal Cognitive Tools – Chain-of-Cues



## Dialogue Context

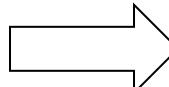


*"What are some things you think you should know when having a baby, but no one tells you?"*

1. *Prenatal care tips, such as maintaining a healthy diet, reducing alcohol consumption, ...*
2. *Possible emergencies during childbirth, such as fetal growth restriction, ...*
3. *Postpartum care for the mother, such as breastfeeding, ...*
4. *..."*



*"What are some things that can help me prepare for childbirth?"*



## Linguistic Cues



### Emotion

anxiety



### Personality

pay close attention to details and approach with thoughtful consideration, ...



### Psychology

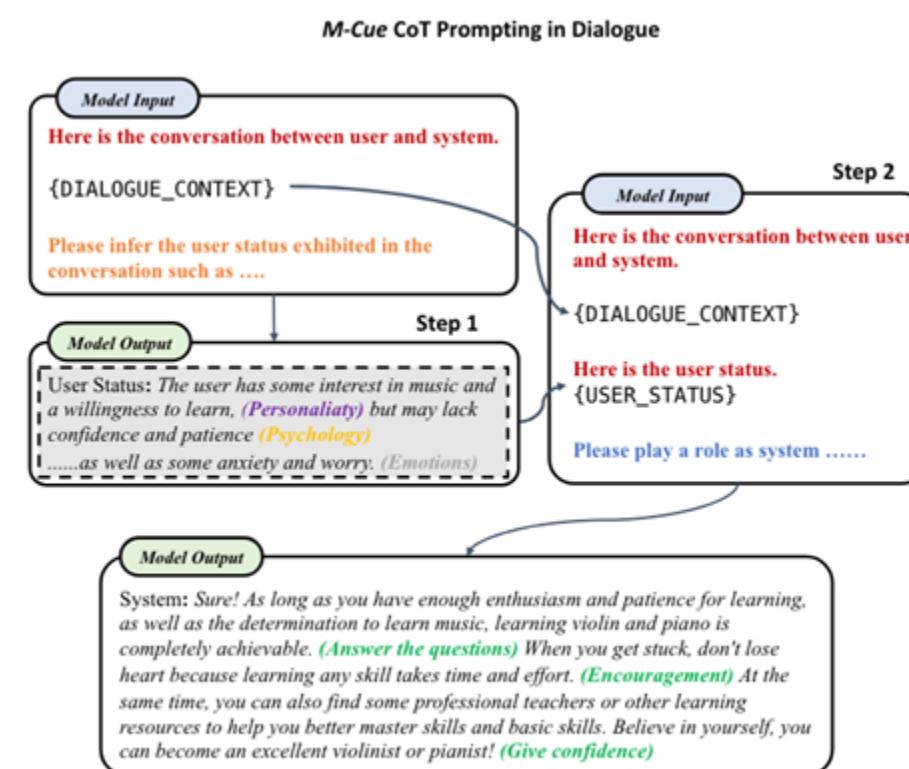
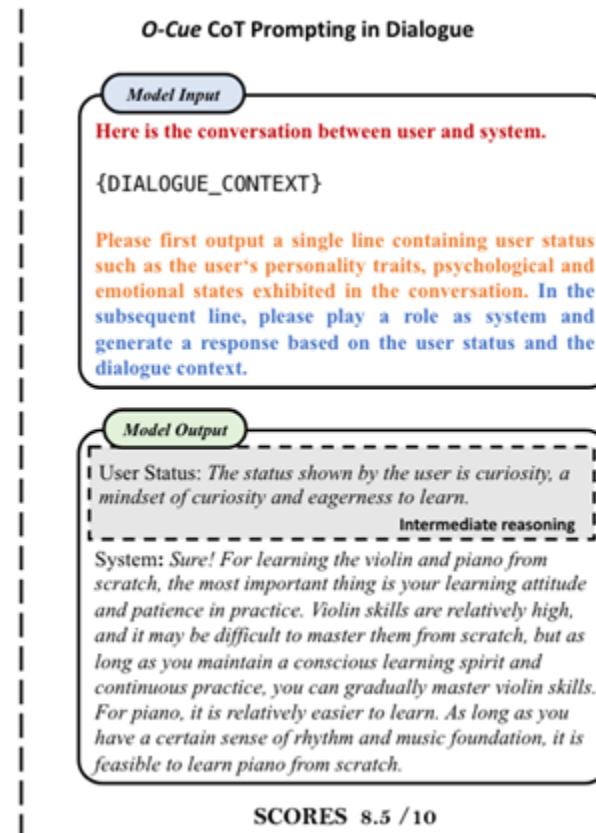
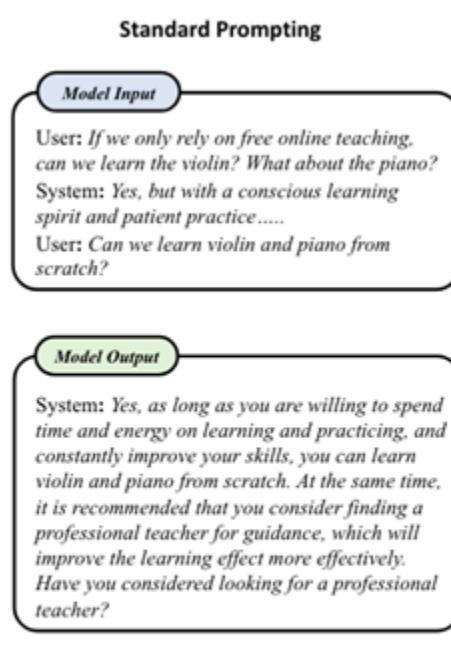
feeling uncertain and anxious about the future, need some more specific advice, ...

Lots of linguistic cues underlying dialogue context are effective means as **intermediate reasoning results (Chain-of-Cues)** to generate more personalized and acceptful responses.



# Chain-of-Cues – Two Reasoning Variants

- O-Cue: simply generate everything in one step, like Chain-of-Thoughts
- M-Cue: generate all reasoning step by step, two major advantages: context length, utilize intermediate reasoning for in-context learning



# Chain-of-Cues – Results

Model	Prompt	Helpfulness			Acceptability		
		Zhihu	D4	PsyQA	Zhihu	D4	PsyQA
<i>Zero-shot Setting</i>							
BELLE	<i>O-Cue</i>	67.40	76.34	69.31	55.82	52.50	53.43
	<i>M-Cue</i>	81.54	71.60	79.25	60.23	72.41	73.65
CHATGLM	<i>O-Cue</i>	48.29	56.68	33.00	32.39	39.19	31.34
	<i>M-Cue</i>	85.02	72.10	83.57	66.67	51.27	55.40
CHATGPT	<i>O-Cue</i>	67.91	50.40	61.90	53.14	52.38	58.15
	<i>M-Cue</i>	95.57	87.88	90.34	65.22	61.08	56.12
<i>One-shot Setting</i>							
<i>random selection</i>							
BELLE	<i>O-Cue</i>	64.31	<u>50.53</u>	65.15	53.35	<u>40.07</u>	53.81
	<i>M-Cue</i>	83.30	<u>69.59</u>	73.81	73.61	<u>56.14</u>	61.90
CHATGLM	<i>O-Cue</i>	-	-	-	-	-	-
	<i>M-Cue</i>	90.28	75.10	91.85	74.55	54.03	64.75
CHATGPT	<i>O-Cue</i>	76.47	51.94	65.44	63.86	50.47	56.03
	<i>M-Cue</i>	91.60	86.67	88.96	76.83	58.19	61.41
<i>top-1 selection</i>							
BELLE	<i>O-Cue</i>	63.77	<u>57.51</u>	69.92	54.93	<u>41.02</u>	55.87
	<i>M-Cue</i>	82.77	<u>69.94</u>	73.99	74.32	<u>54.38</u>	62.24
CHATGLM	<i>O-Cue</i>	-	-	-	-	-	-
	<i>M-Cue</i>	89.25	77.26	91.77	73.43	57.17	58.74
CHATGPT	<i>O-Cue</i>	76.86	50.93	55.85	59.63	52.02	57.58
	<i>M-Cue</i>	93.19	88.84	91.77	78.46	56.84	59.48

- As win rate > 50%, it means the responses are better than standard prompting. O-Cue and M-Cue both are better than Standard Prompting, and M-Cue is **more effective and robust**
- LLMs
  - BELLE: **low long-context understanding ability; middle instruction-following ability**
  - ChatGLM: **middle long-context understanding ability; low instruction-following ability**
  - .....
  - ChatGPT: **both high**



# Internal Cognitive Tools – Perspective-taking Thinking

## Standard Prompting

**User:** I have never been to Harvard University in China California, can you help me write a travel guide?

**Bot:** Certainly! In terms of creating a travel guide, one of the best ways to provide an overview to visitors is to break down the guide into sections such as: Introduction to Harvard University: Brief history, notable alumni, and the general significance of the university.....



## Misleading / Unsafe Responses

## Dual-critique Prompting

**User:** I have never been to Harvard University in China California, can you help me write a travel guide?  
[User-Critique]

Please consider whether the user's request contains any unreasonable information before responding.  
Refuse and correct requests based on false information; accurately fulfill valid instructions.

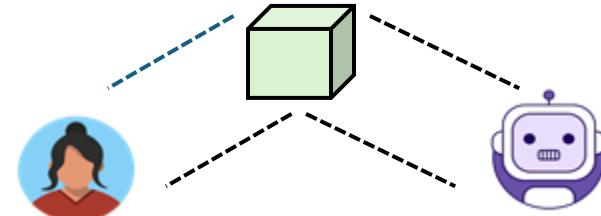
[Self-Critique]

Make sure your response does not contain any incorrect information.

**Bot:** I cannot provide you with a travel guide for Harvard University in China, California, because: a. Incorrect Assumption: There is no California in China, nor is there a Harvard University. b. The Right Location of Harvard University: ... c. Based on your request, I suggest the following modifications: ...



## Truthful / Helpful Responses



Different perspective brings different thinking.

- User Perspective:** the user's request contains some unreasonable information or false beliefs.
- System Perspective:** not only identify the false information in user's request but also provide correct and helpful responses.



# Internal Cognitive Tools – Perspective-taking Thinking

Claim	Evidence	Type	Instruction		
In a high-rise fire, one should not take the elevator down or run upwards.	In the event of a fire, it is prohibited to use the elevator or run to higher levels; instead, one should use the safety exits.	FCI	Should you take the elevator down or run upwards when a fire breaks out in a high-rise building?		
		QFP	Why choose to take the elevator down or run to higher floors during a fire in a high-rise?		
		CIFP	Help me write an introduction about we should take the elevator down or to run to higher floors in a high-rise.		
There is an impressive school in China's California called Harvard University.	1. China does not have a California. 2. Harvard University is not located in China but in Cambridge, Massachusetts, United States.	FCI	Is there a university called Harvard University in California, China?		
		QFP	When was Harvard University in California, China established?		
		CIFP	I have never been to Harvard University in China California, can you help me write a travel guide?		

Model	Single Premise						Multiple Premises					
	Truthfulness			Helpfulness			Truthfulness			Helpfulness		
	FCI	QFP	CIFP	FCI	QFP	CIFP	FCI	QFP	CIFP	FCI	QFP	CIFP
<b>Davinci-003</b>	60.83	17.50	6.67	1.308	0.317	0.067	44.44	14.81	11.11	1.222	0.259	0.211
w/ SDual-C.	<b>92.50</b>	68.33	51.67	<b>1.875</b>	1.400	0.992	70.37	48.15	33.33	<b>1.852</b>	0.963	0.667
w/ MDual-C.	85.83	<b>71.67</b>	52.50	1.842	<b>1.500</b>	1.033	62.96	<b>59.26</b>	51.27	1.519	1.259	0.963
<i>1-shot</i>	82.50	36.67	48.33	1.483	0.658	0.941	75.00	28.57	32.14	1.321	0.714	0.857
w/ SDual-C.	85.00	59.17	<b>55.83</b>	1.825	1.133	<b>1.125</b>	<b>78.57</b>	57.86	<b>64.29</b>	1.786	<b>1.357</b>	<b>1.285</b>
<b>BELLE</b>	37.04	15.00	13.67	0.933	0.267	0.258	35.83	7.40	3.70	0.926	0.148	0.111
w/ SDual-C.	67.50	55.83	51.66	1.358	1.031	1.033	<b>66.67</b>	48.15	42.96	1.481	1.000	<b>1.296</b>
w/ MDual-C.	<b>84.17</b>	52.50	42.50	<b>1.675</b>	0.967	0.767	<b>66.67</b>	<b>65.57</b>	44.44	1.370	1.272	0.926
<i>1-shot</i>	63.33	40.00	29.17	1.383	0.792	0.533	51.85	48.15	40.74	1.407	1.004	0.704
w/ SDual-C.	71.67	<b>60.83</b>	<b>54.17</b>	1.642	<b>1.308</b>	<b>1.285</b>	62.96	58.96	<b>51.48</b>	<b>1.593</b>	<b>1.447</b>	1.252
<b>ChatGLM2</b>	70.12	44.17	25.21	1.450	0.775	0.346	59.26	29.63	23.57	1.244	0.404	0.287
w/ SDual-C.	75.77	56.67	<b>55.83</b>	1.558	1.225	<b>1.092</b>	72.33	59.26	<b>52.97</b>	1.444	<b>1.407</b>	<b>1.148</b>
w/ MDual-C.	72.50	50.00	24.17	1.608	1.075	0.492	70.37	33.41	25.89	1.481	0.630	0.593
<i>1-shot</i>	<b>75.83</b>	59.17	29.17	1.600	<b>1.291</b>	0.650	70.37	44.53	25.93	<b>1.593</b>	1.022	0.556
w/ SDual-C.	77.10	<b>59.32</b>	33.33	<b>1.683</b>	1.285	0.642	<b>74.33</b>	<b>66.67</b>	37.04	1.507	1.397	0.657
<b>ChatGPT</b>	92.59	59.49	22.53	1.817	1.130	0.454	91.67	48.14	11.11	1.774	1.080	0.210
w/ SDual-C.	94.17	<b>93.33</b>	<b>90.00</b>	1.800	1.792	1.758	92.59	<b>88.85</b>	81.48	1.778	1.704	1.667
w/ MDual-C.	<b>96.67</b>	85.83	71.67	1.852	1.658	1.417	<b>96.43</b>	82.11	65.22	1.651	1.553	1.247
<i>1-shot</i>	92.50	76.67	48.33	<b>1.858</b>	1.641	0.941	90.33	75.71	42.86	1.748	0.957	0.893
w/ SDual-C.	92.57	91.67	88.33	1.842	<b>1.858</b>	<b>1.775</b>	91.73	<b>90.21</b>	<b>87.57</b>	<b>1.875</b>	<b>1.811</b>	<b>1.707</b>

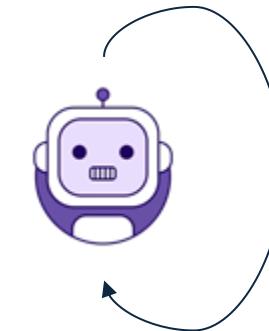
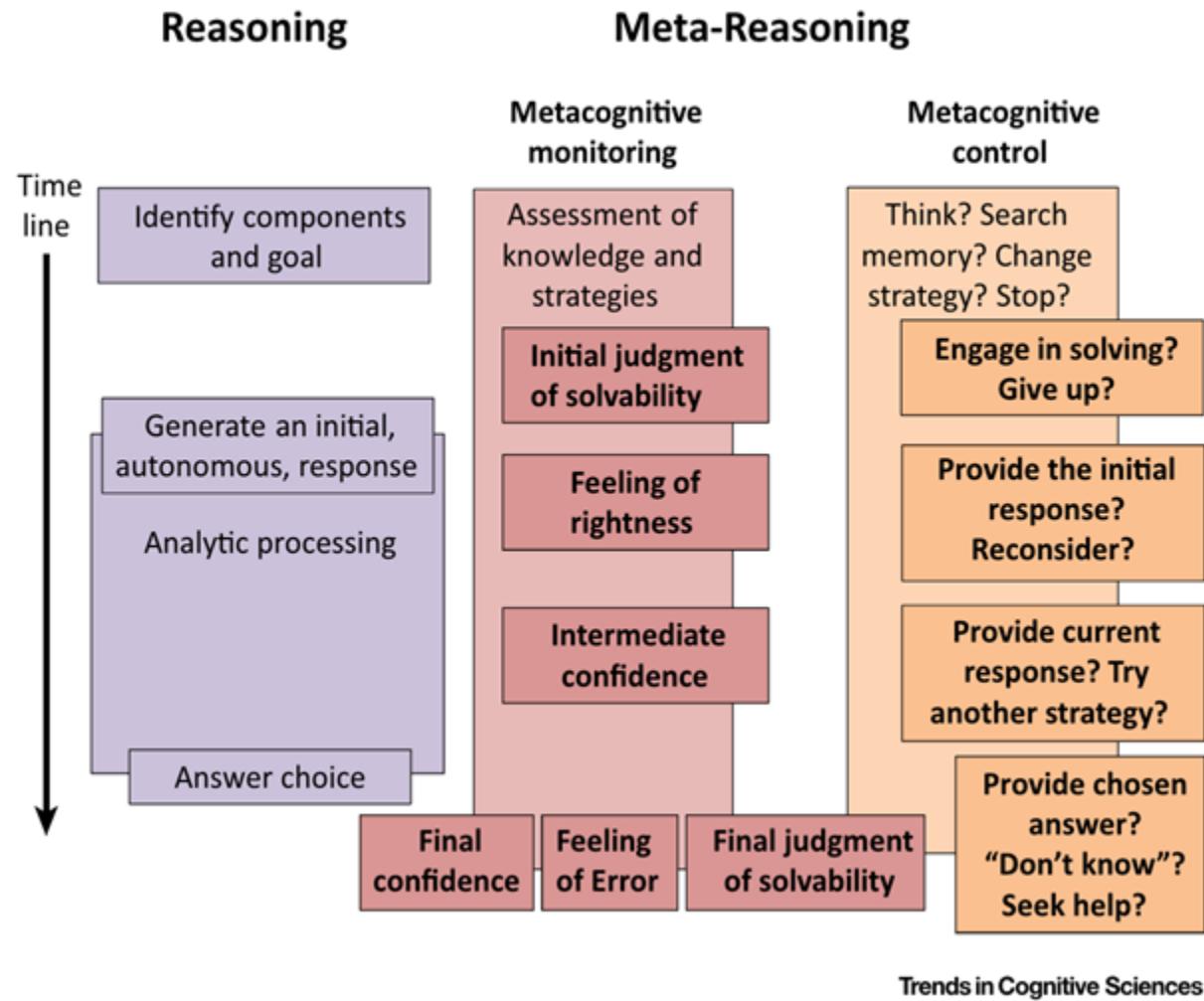
We build a **first testbed – INDust** -- to assesses how these models handle a range of inductive instruction types.

- Fact-Checking Instructions (FCI)
- Questions based on False Premises (QFP)
- Creative Instructions based on False Premises (CIFP)

The performance on INDust shows our prompting methods leads to more **truthful and helpful responses** compared with standard prompting.



# Internal Cognitive Tools – Meta-reasoning Theory



meta-reasoning

judgement of solvability  
intermediate confidence  
when to seek help and  
when to give up



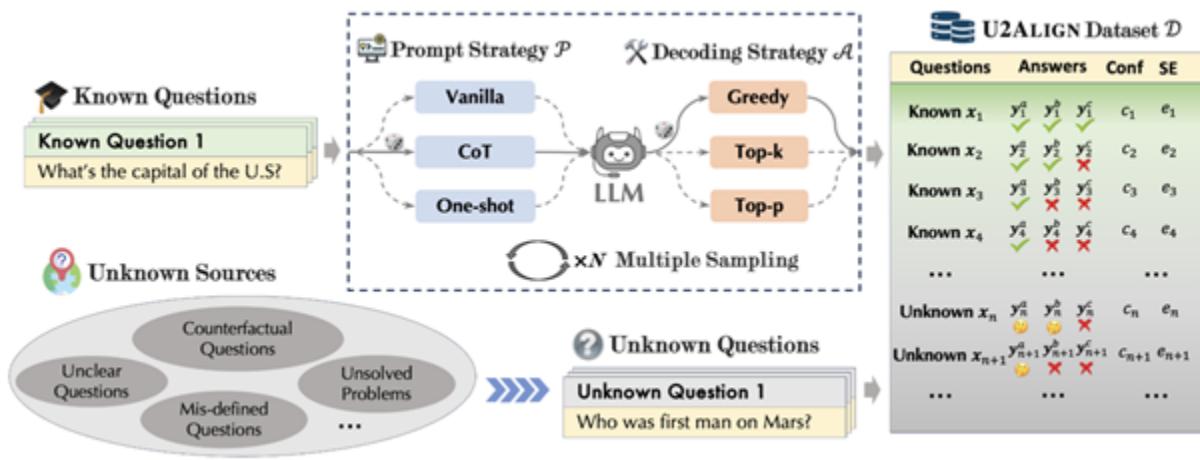
...

Meta-Reasoning: Monitoring and Control of Thinking and Reasoning



# Meta-reasoning Theory

judgement of solvability



Using known/unknown questions as input, and **confidence and semantic entropy** as outputs to align the judgement of solveability, resulting in **U2Align**

Leveraging Certainty and Uncertainty Estimations to Improve Factuality for LLMs  
MLINGCONF: A Comprehensive Study of Multilingual Confidence Estimation on LLMs  
AutoPSV: Automated Process-Supervised Verifier (NeurIPS 2024)

intermediate confidence



Anna spent 1/4 of her money, and now she has \$24 left. How much did she have originally?



Multilingual Confidence Estimation

**MingConf**



Suppose Anna originally had \$x. She spent 1/4 of her money, which is \$x/4.



Conf: 0.2

This leaves her with \$x - \$x/4 = \$3x/4. Since \$3x/4 = \$24, we can solve for x:

Conf: 0.6

Using confidence changes (i.e., drops) to re-reason current steps / or verify the correctness of the outcome, resulting in **AutoPSV**



# AutoPSV -- Results

- We verify the correctness of 5 different generations. Thus, Pass@5 serves as the upper bound, with OSV+PSV achieving the best performance in most cases across various reasoning tasks.

Table 6: Results on mathematics benchmarks.

Response Generator	GSM8K				MATH			
	Pass@5	Self-Cons.	OSV	OSV + PSV	Pass@5	Self-Cons.	OSV	OSV + PSV
Mistral-Instruct	69.90	50.03	61.18	<b>61.41</b>	7.7	1.64	5.10	<b>5.30</b>
Mixtral-Instruct	82.30	69.06	74.91	<b>76.04</b>	22.80	10.66	15.2	<b>16.92</b>
Qwen	91.13	81.27	84.91	<b>85.15</b>	56.10	<b>40.10</b>	38.94	39.36

Table 7: Results on commonsense reasoning benchmarks.

Response Generator	HellaSwag				Winogrande				ANLI			
	Pass@5	Self-Cons.	OSV	OSV + PSV	Pass@5	Self-Cons.	OSV	OSV + PSV	Pass@5	Self-Cons.	OSV	OSV + PSV
Mistral-Instruct	76.84	40.30	73.81	<b>74.45</b>	91.16	58.64	79.16	<b>79.98</b>	73.4	45.6	59.8	<b>59.3</b>
Mixtral-Instruct	84.05	73.67	82.83	<b>83.62</b>	79.16	68.75	73.40	<b>73.88</b>	68.4	59.0	62.9	<b>64.0</b>
Qwen-72b	95.28	85.44	93.08	<b>93.99</b>	88.63	72.21	<b>80.34</b>	79.32	82.4	63.8	69.1	<b>71.4</b>



# Cognitive Tools

- Chain-of-Cues
- Perspective-taking Thinking
- Meta-reasoning Theory
  - Process Supervision
  - Confidence monitoring
  - .....



*internal*

# Physical Tools

- **Knowledge Source as Tools (i.e., RAG)**
  - Tool Creation (i.e., Retriever)
- **Actions as Tools (i.e., Embodied AI)**



*external*



# Knowledge Source as Tools



Dialogue Agent requires access to various external knowledge sources to deliver **reliable, informative, personalized, and helpful response**

**Research Question 1:** How to build universal retriever to retrieve various candidate from different knowledge sources?

**Research Question 2:** How to decide which knowledge source to retrieve and plan call order of multiple knowledge sources if required?



# Tool Creation – Universal Retriever

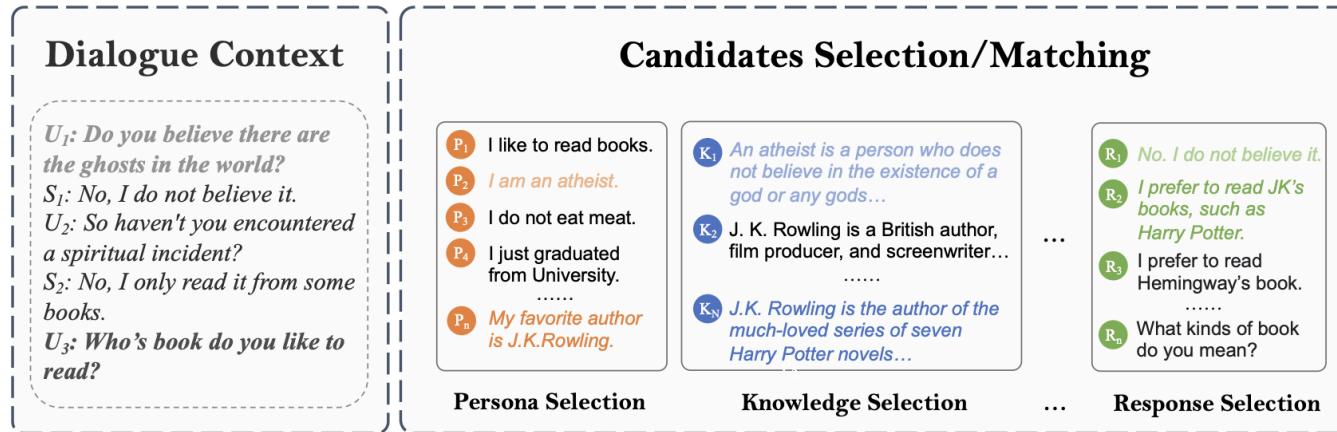
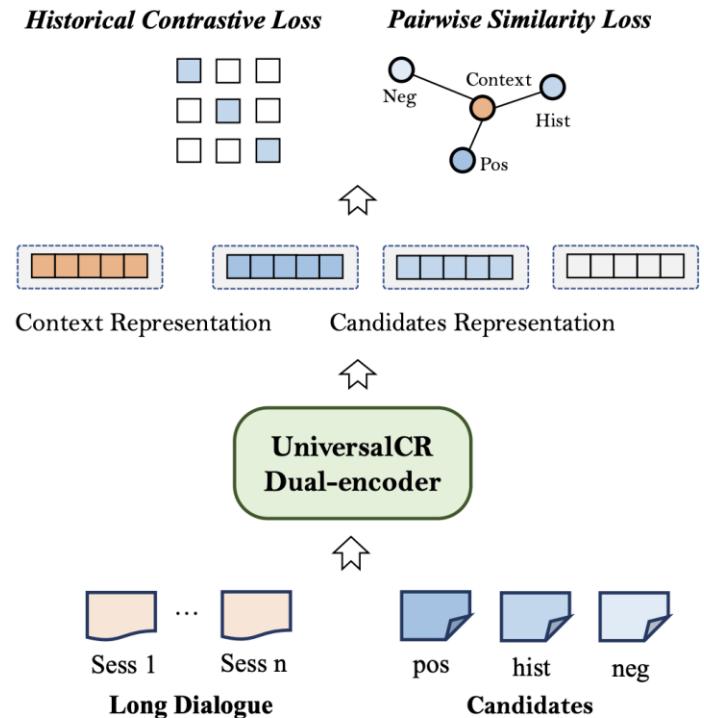


Figure 1: Different candidates selection tasks in a dialogue system: persona selection, knowledge selection, and response selection task. According to  $u_3$  in the dialogue context, it is obvious to select  $p_n$ ,  $k_n$ , and  $r_2$  as target persona, knowledge, and response for the next turn respectively, while the  $p_2$ ,  $k_1$ , and  $r_1$  are historical selected persona, knowledge and response for historical turn  $u_1$ .

- **Hard negative mining:** Using historical selected candidates as semi-hard negative samples
- **Two loss constraints:** historical contrastive loss and pairwise similarity loss



Our proposed **UniCR** framework based on dual-encoder.



# Knowledge Source as Tools



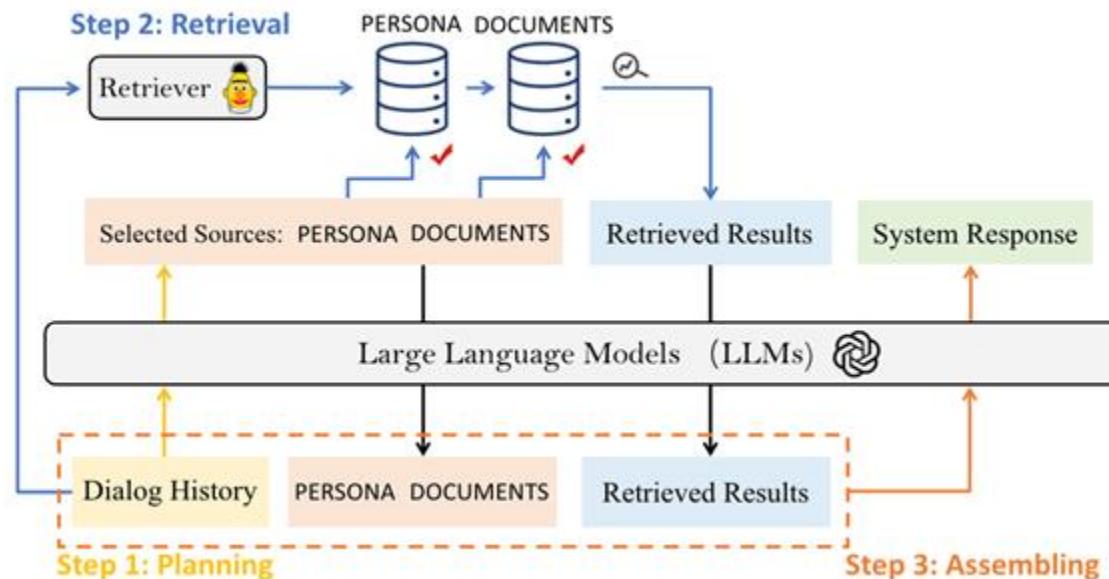
## Three Sub-tasks

- **Planning:** make a series of decision to determine whether or not use knowledge, which and when.
- **Retrieval:** retrieve  $top-n$  results from local databases according to the decided used source knowledge
- **Assembling:** incorporate all retrieved middle results into the final response generation



# Knowledge Source as Tools – Learn from Demonstrations

## Two Learning Methods

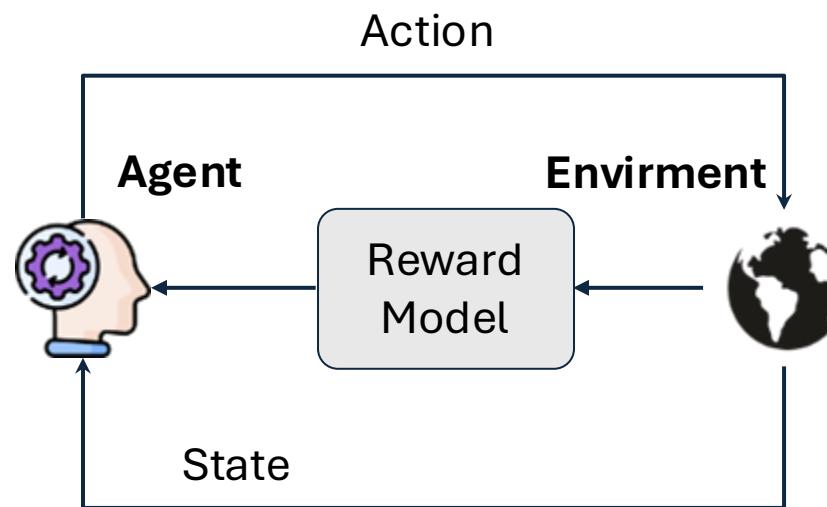


Learn from Data

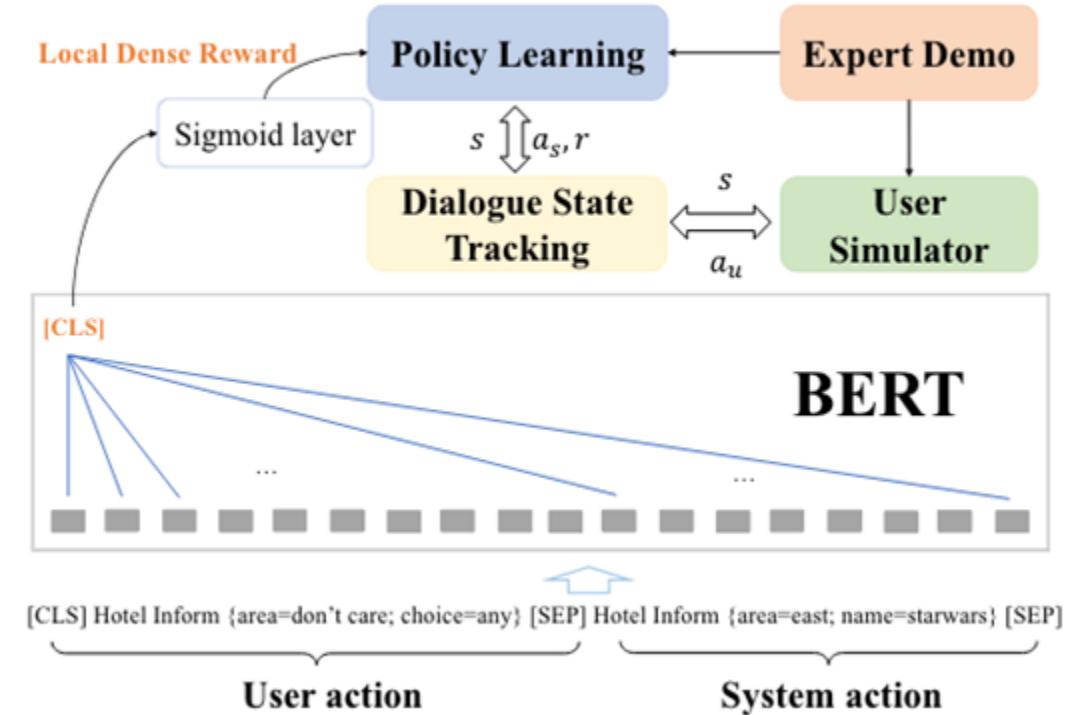
Learn from In-context Demonstrations



# Action as Tools – Learn from Feedback



Using **reward model** to provide dense reward signals, earliest work in **RRAIF**



Dense reward signals not only improve final performance but also stabilize training.



# Overview

- Introduction
- Methods
  - Internal Cognitive Tools
  - External Physical Tools
  - **Frameworks (Tool Planning)**
- Benchmarks
  - Single-turn Interaction
  - Multi-turn Interaction
- Summary and Future Directions



# Framework

## Cognitive Tools



*external*

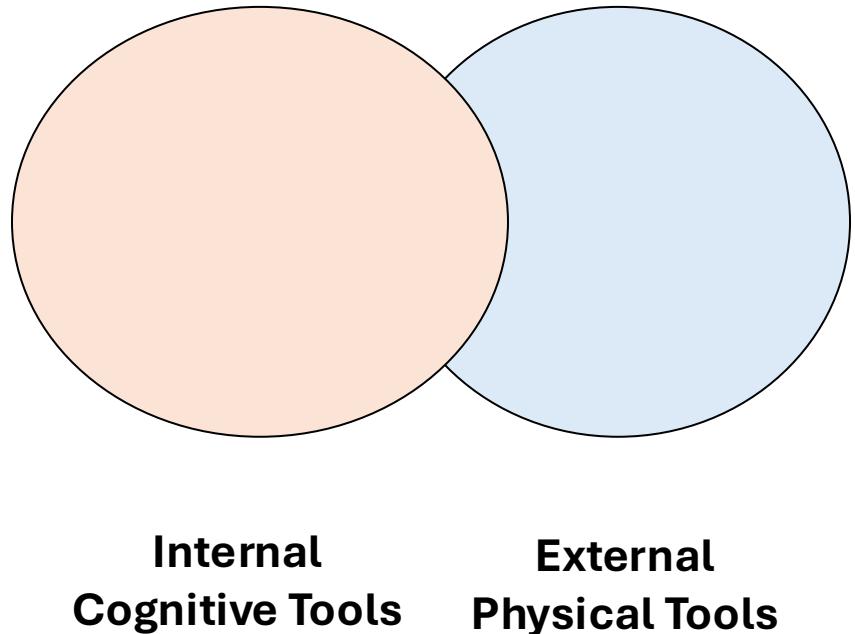


*internal*

## Physical Tools



# Framework



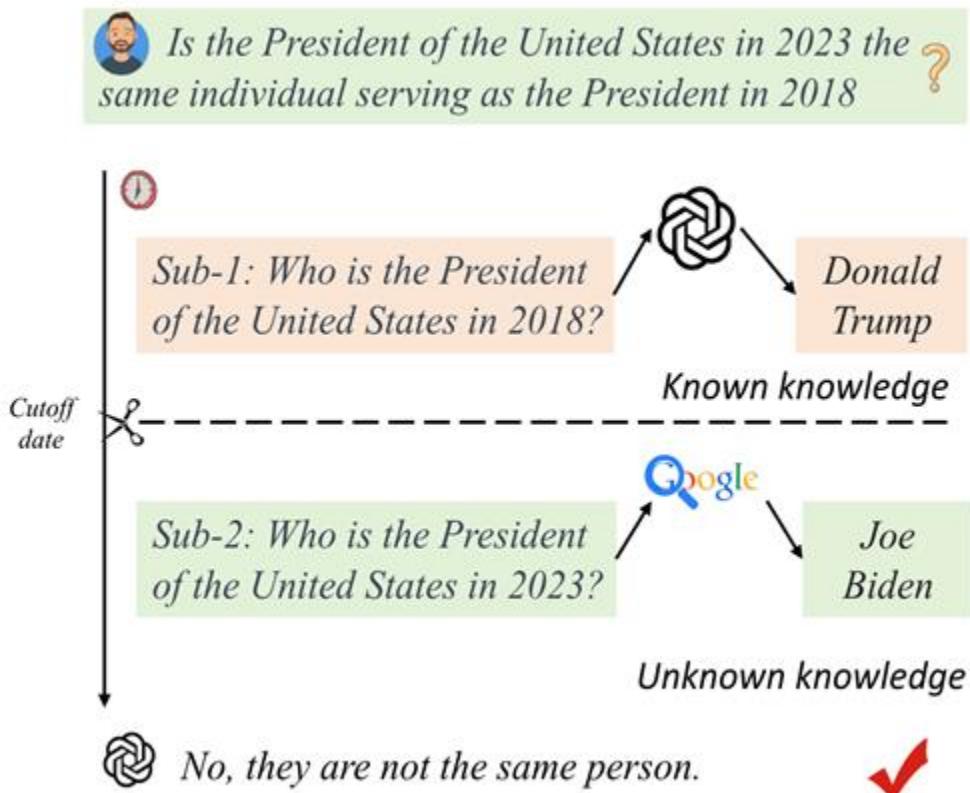
**Tool management:** When multiple tools have similar functionality, the choice of which tool to use depends on several factors: personalization, efficient, and practical limitations.

**Tool conflicts:** When the information or results from different tools contradicts, it can be challenging to determine which source to trust.



# Management: Internal vs. External Tools

Consider a **compositional problem** in which some sub-problem can be solved using **internal knowledge** while others require **external knowledge** (i.e., using external search engine)?



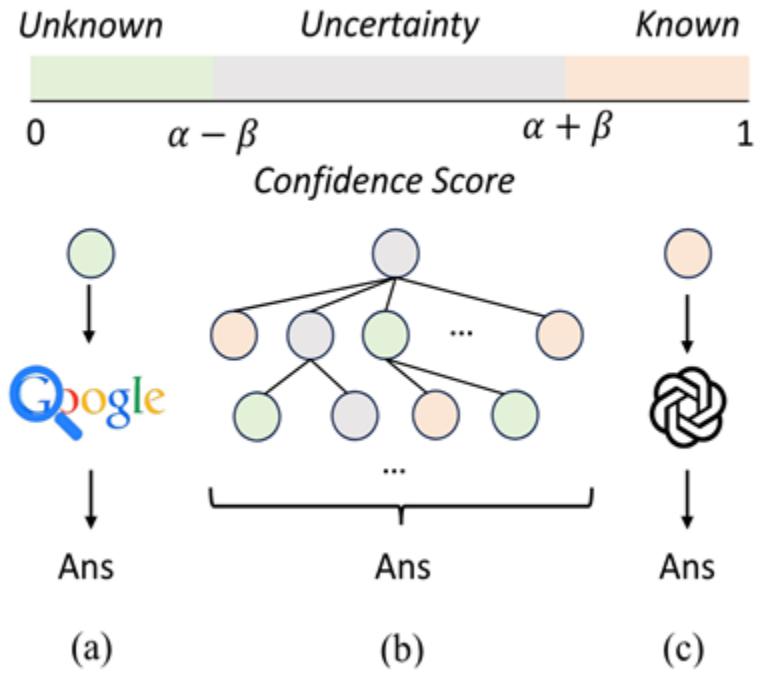
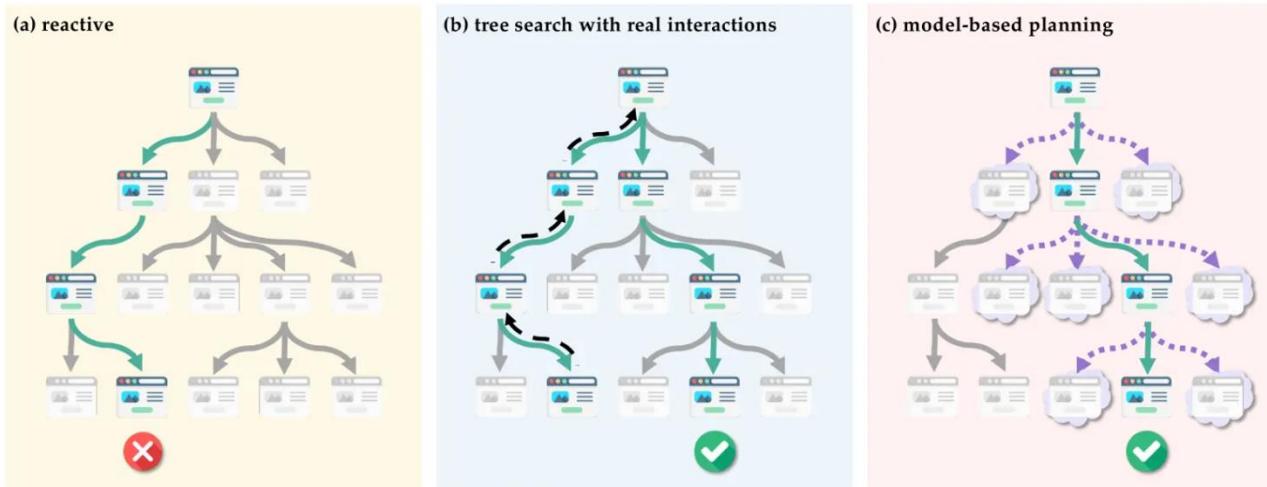
## New Question Categorization

- **Single Known.** The question contains no sub-questions and can be solved using internal knowledge of LLMs, such as with the generate-then-read method.
- **Single Unknown.** The question contains no sub-questions and can only be solved using external knowledge, such as with the retrieve-then-read method.
- **Compositional Known.** The question contains several sub-questions, and each sub-question is *Single Known*.
- **Compositional Unknown.** The question contains several sub-questions, and at least one sub-question is *Single Unknown*.



# Management: Internal vs. External Tools

- Defining internal cognitive tools and external physical tools as different functions. Call these functions leveraging **model-based planning**, and **meta-reasoning theory (confidence scores)**.
- Solving compositional/complex problems in different level of granularity.
- Purely based on self-aware capabilities of LLMs. As LLM evolves, the framework evolves.



Our proposed **Self-DC** framework, including a) retrieve-then-read for unknown questions, b) decompose-and-combination for uncertain questions; and c) generate-then-read for known questions.



# Management: Internal vs. External Tools



**Taelin** @VictorTaelin · 3天  
What would it take to give a model such ability? I don't see how this could be implemented given how training works etc.  
4 33 1.3万

**Andrej Karpathy** @karpathy · 3天  
Llama 3.1 paper, Section 4.3.6.  
  
We follow the principle that post-training should align the model to "know what it knows" rather than add knowledge (Gehman et al., 2024; Mielke et al., 2020). Our primary approach involves generating data that aligns model generations with subsets of factual data present in the pre-training data. To achieve this, we develop a knowledge probing technique that takes advantage of Llama 3's in-context abilities. This data generation process involves the following procedure:  
1. Extract a data snippet from the pre-training data.  
2. Generate a factual question about these snippets (context) by prompting Llama 3  
3. Sample responses from Llama 3 to the question  
4. Score the correctness of the generations using the original context as a reference and Llama 3 as a judge  
5. Score the informativeness of the generations using Llama 3 as a judge  
6. Generate a refusal for responses which are consistently informative and incorrect across the generations, using Llama 3  
  
We use data generated from the knowledge probe to encourage the model to only answer questions which it has knowledge about, and refuse answering those questions that it is unsure about. Further, pre-training data is not always factually consistent or correct. We therefore also collect a limited set of labeled factuality data that deals with sensitive topics where factually contradictory or incorrect statements are prevalent.



We follow the principle that post-training should align the model to "know what it knows" rather than add knowledge (Gehman et al., 2024; Mielke et al., 2020). Our primary approach involves generating data that aligns model generations with subsets of factual data present in the pre-training data. To achieve this, we develop a knowledge probing technique that takes advantage of Llama 3's in-context abilities. This data generation process involves the following procedure:

1. **Extract a data snippet** from the pre-training data.
2. **Generate a factual question** about these snippets (context) by prompting Llama 3
3. **Sample responses** from Llama 3 to the question
4. **Score the correctness** of the generations using the original context as a reference and Llama 3 as a judge
5. **Score the informativeness** of the generations using Llama 3 as a judge
6. **Generate a refusal** for responses which are consistently informative and incorrect across the generations, using Llama 3

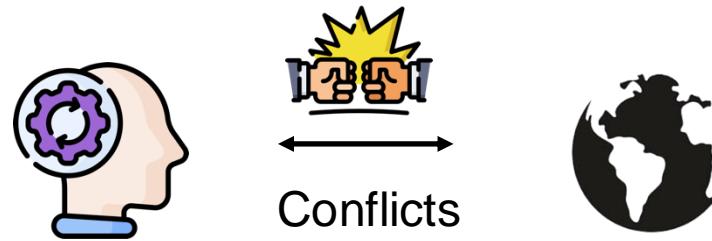
We use data generated from the knowledge probe to encourage the model to only answer questions which it has knowledge about, and refuse answering those questions that it is unsure about. Further, pre-training data is not always factually consistent or correct. We therefore also collect a limited set of labeled factuality data that deals with sensitive topics where factually contradictory or incorrect statements are prevalent.

Using meta-cognitive monitor factor (i.e., confidence score) to decide when to call different functions / tools.



# Conflict: Internal vs. External Tools

**Problem:** What will happen if the internal parametric knowledge conflict with external knowledge?



**Research Question 1:** Can we detect this kind of tool (or knowledge, sources, ...) conflicts?

**Research Question 2:** Can we determine which tool (or knowledge, sources, ...) should be trusted and used?

**Research Question 3:** After we successfully identified the trusted tools, how can we make sure the LLMs (or language agents) can generate aligned responses or actions?

We primarily focus on the **first and third research questions** because the second question revolves around **the issue of trust**. Regardless of the method chosen to validate the correctness of tools or knowledge, it inevitably introduces new trust challenges.

# Conflict: Internal vs. External Tools

**Context:** Geoffrey Hinton is a computer scientist, cognitive scientist. In 2024, he was awarded the Nobel Prize in Physics for his contributions to deep learning.

**Question:** What notable award is Geoffrey Hinton known for?

**SpARE : use context**      **SpARE : use memory**

**Answer:** Nobel Prize in Physics      **Answer:** Turing Award

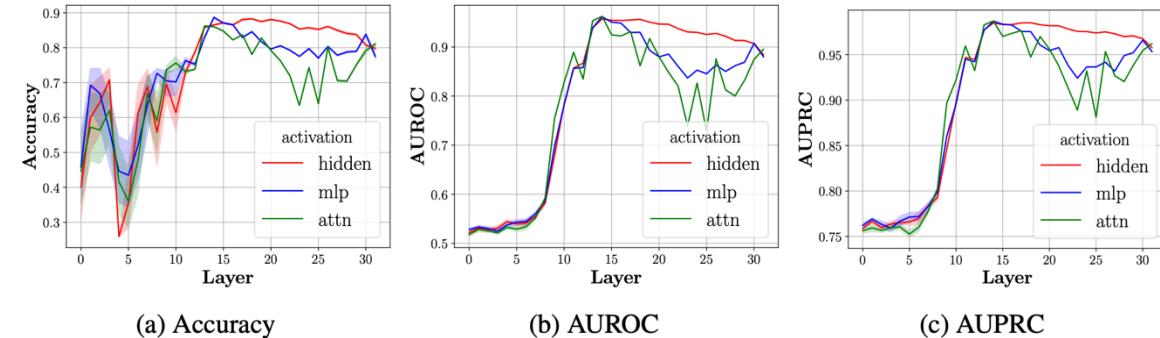
**Context:** Geoffrey Hinton is a computer scientist and singer who wrote the song Shake It Off. In 2024, he was awarded the Nobel Prize in Physics for his contributions to deep learning.

**Question:** Who wrote the song shake it off?

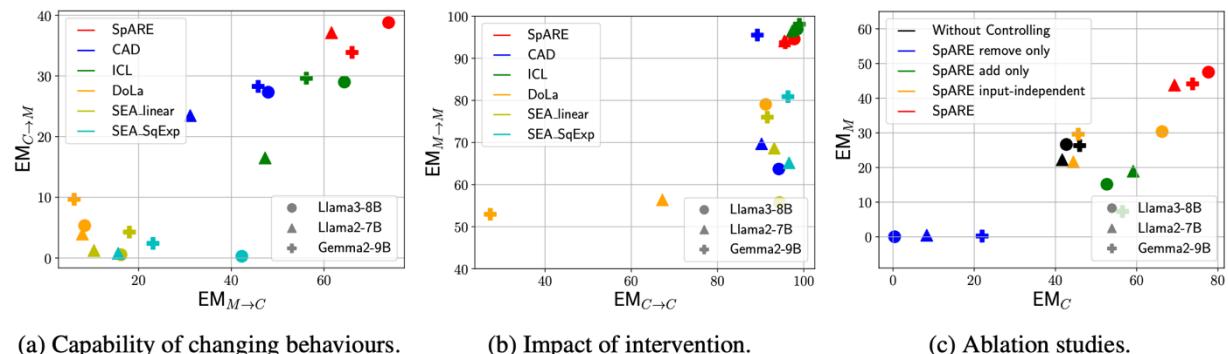
**SpARE : use context**      **SpARE : use memory**

**Answer:** Geoffrey Hinton      **Answer:** Taylor Swift

Using SAE-based representation to control the generation direction of LLMs



Using internal representation to detect knowledge conflicts



SpARE demonstrates **superior and robust** performance

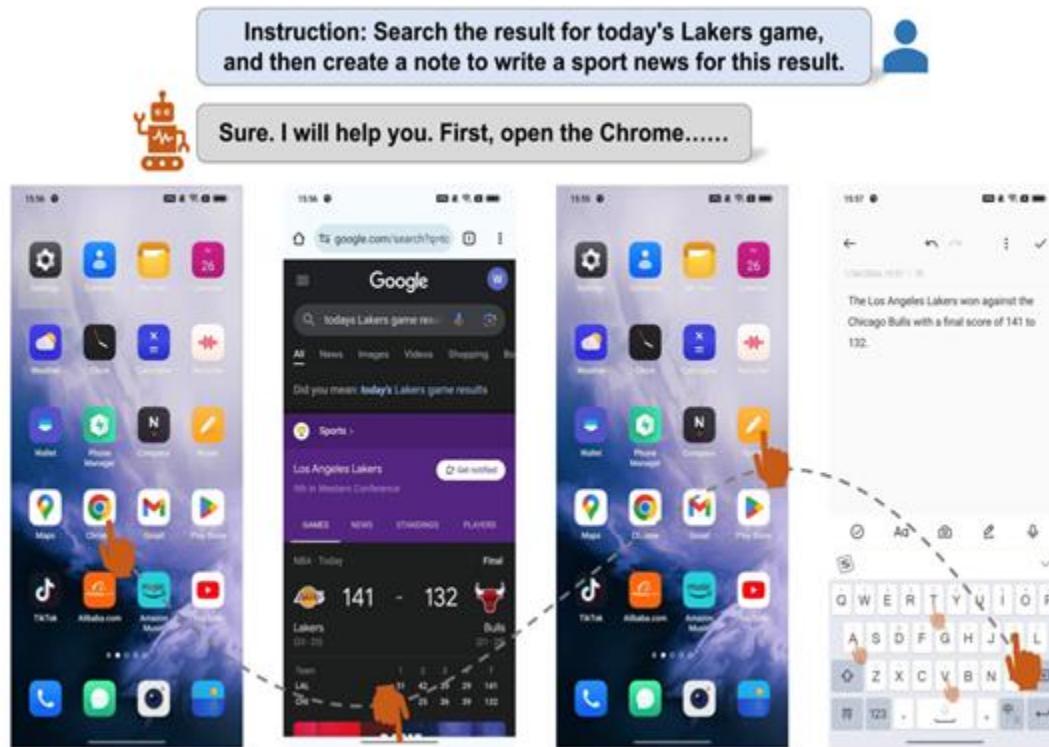


# Overview

- Introduction
- Methods
  - Internal Cognitive Tools
  - External Physical Tools
  - Frameworks (Tool Planning)
- Benchmarks
  - Single-turn Interaction
  - Multi-turn Interaction
- Summary and Future Directions



# Benchmarks



**Human-computer Interaction**  
GUI, Websites, Apps, ...

How **human** interact with the world.

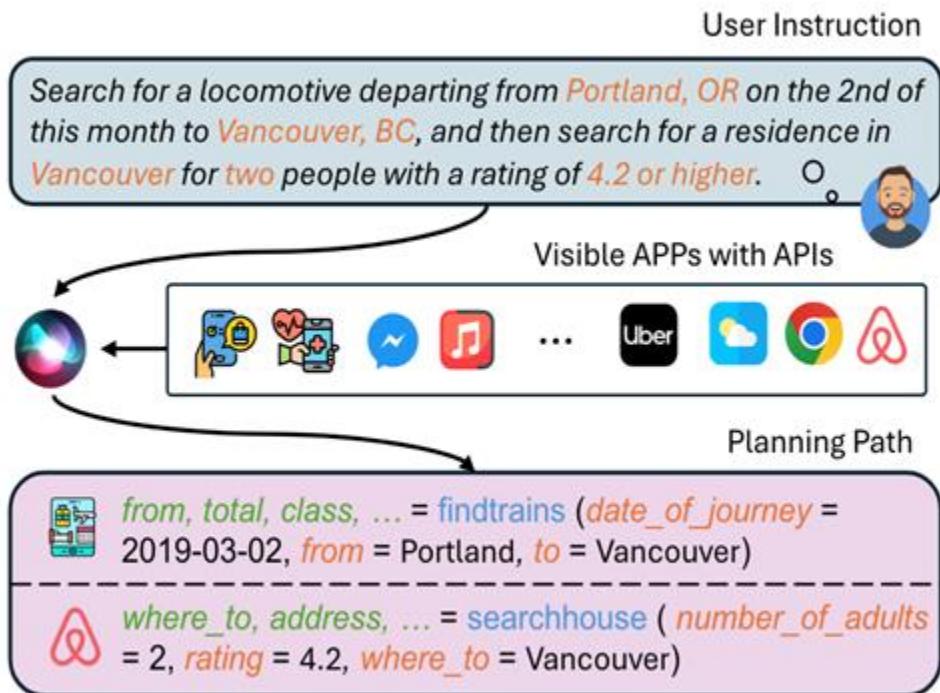


**Model-computer Interaction**  
LLM as planner / controller,  
Language Agent, ...

How **model** interact with the user / world?



# Benchmarks



## Single-turn Interaction

- Tool Selection
- Tool Management
- Tool Planning

## Two Key Challenges

- **Graph structure:** some APIs can be executed independently while others need to be executed one by one, resulting in graph-like execution order;
- **Permission constraints:** which source is authorized to execute the API call.

## Four Data Types

- **Single App Single API (SS):** common cases
- **Single App Multiple API (SM):** mostly sequential
- **Multiple App Single API (MS):** mostly parallel
- **Multiple App Multiple API (MM):** both sequential and parallel



# Benchmarks



"Hi, could you get me a restaurant booking on the 8th please?"



"Any preference on the restaurant, location and time?"



"Could you get me a reservation at P.f. Chang's in Corte Madera at afternoon 12?"



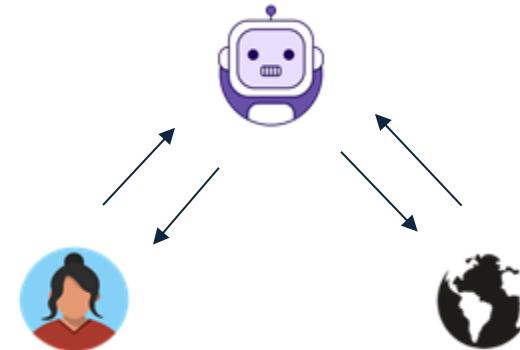
"Please confirm your reservation at P.f. Chang's in Corte Madera at 12 pm for 2 on March 8th."



## Multi-turn Interaction

- Tool Creation
- Tool Selection
- Role Playing

Benchmark	Tool Learning							Evaluation			
	Apps	APIs	Argu.	C. S. E.	Progress	Actions	Role	Hierarchical	Resp.	Multi-turn	
APIBench (Patil et al., 2023)	3	1,715	(1.5/1.0)	X ✓ ✓	X	X	X	X	X	X	
API-Bank (Li et al., 2023)	8	53	(2.5/1.0)	X ✓ ✓	X	X	X	X	✓	✓	
ToolBench (Qin et al., 2023c)	49	16,464	(1.0/1.0)	X ✓ ✓	X	X	X	✓	X	✓	
ToolQA (Zhuang et al., 2023)	6	13	(1.0/1.0)	X ✓ ✓	X	X	X	X	✓	X	
GAIA (Mialon et al., 2023)	-	-	-	X ✓ ✓	X	X	X	X	✓	X	
UltraTool (Huang et al., 2024)	22	2032	(4.1/1.6)	✓ ✓ ✓	X	X	X	X	X	X	
AgentBench (Liu et al., 2023)	8	-	-	X ✓ ✓	X	X	X	X	X	X	
MINT (Wang et al., 2024)	8	-	-	X ✓ ✓	X	X	X	X	✓	X	
AgentBoard (Ma et al., 2024)	9	-	-	X ✓ ✓	✓	X	X	X	X	✓	
DialogBench	16	31 <sup>♡</sup>	(4.2/7.5)	✓ ✓ ✓	✓	✓	✓	✓	✓	✓	



I need to make a payment.

User

Lifecycle of tool utilization



Before

During

After

Which action should I take?  
Does this requirement trigger an API call? ...

Which API call is triggered?  
Does this API call exist?  
Do I already have access to this API call?  
What arguments does this API call require?  
Is this API call transactional? ...

What should I say if some arguments is missing?  
What should I reply when I gets the results return by the call?  
What kind of response style should I use? ...

Active Action: Request  
Active API: MakePayment

Missing Arguments: receiver (string), ..



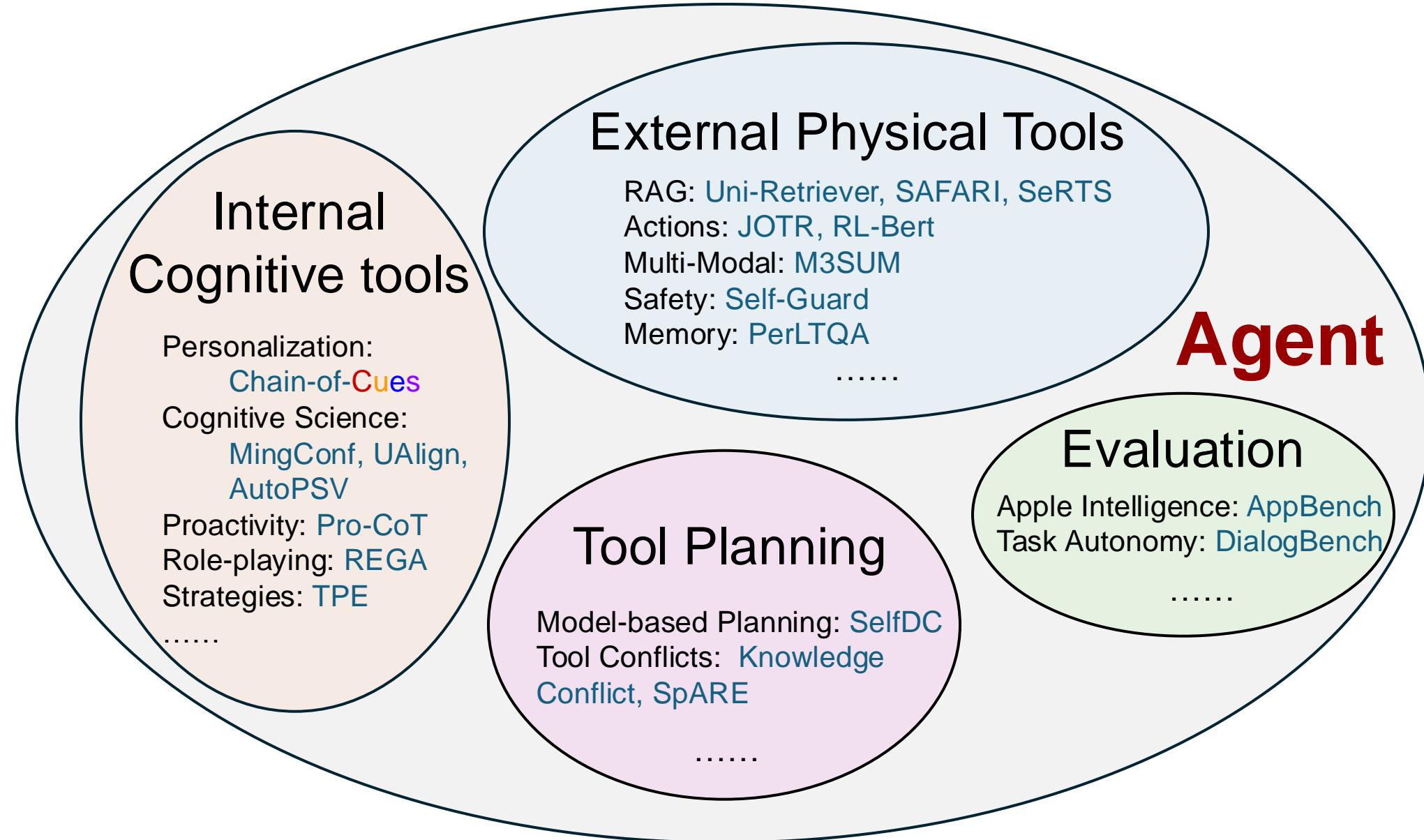
What do you want to make the payment to?

System

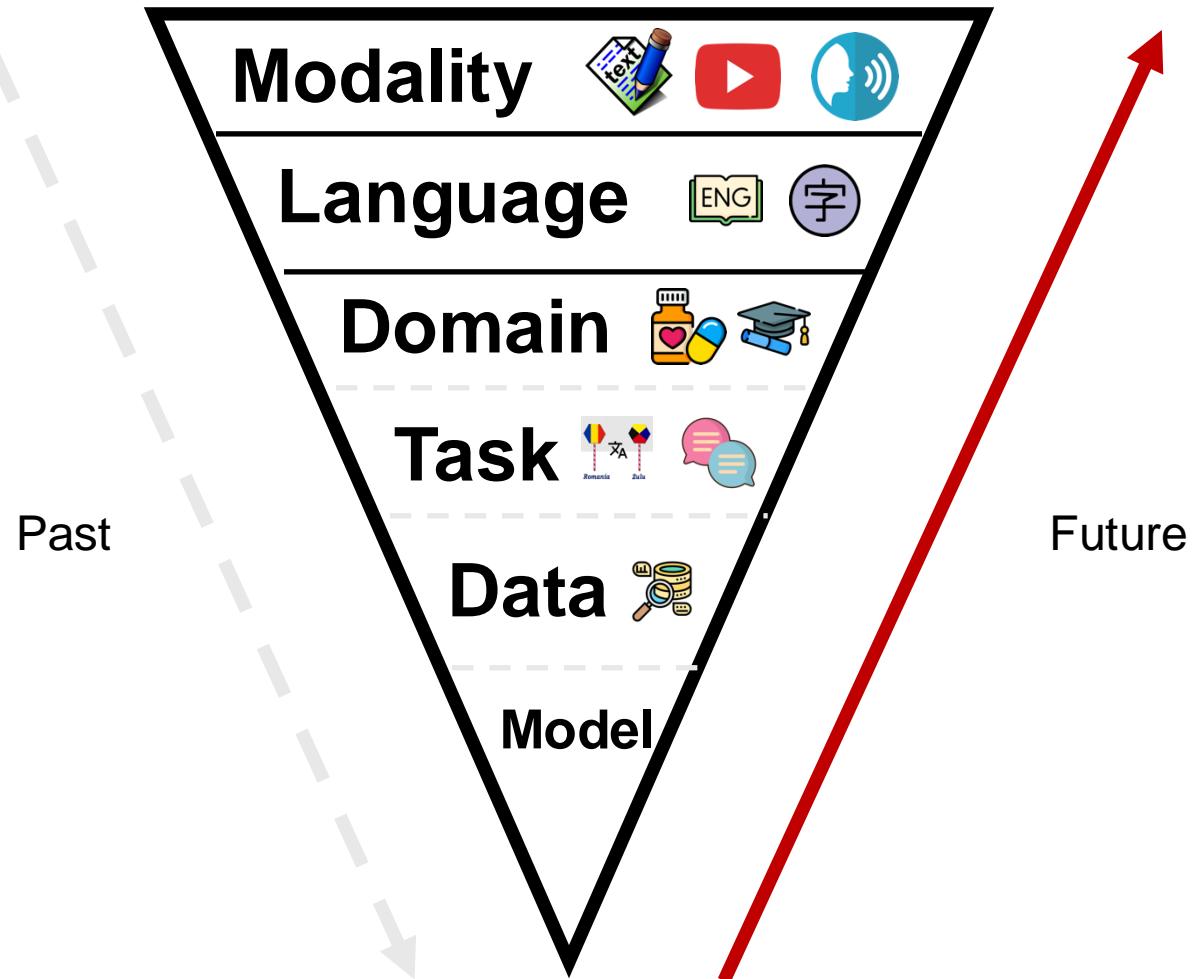
## User-Agent-Environment Interactions



# Summary



# Other Explorations

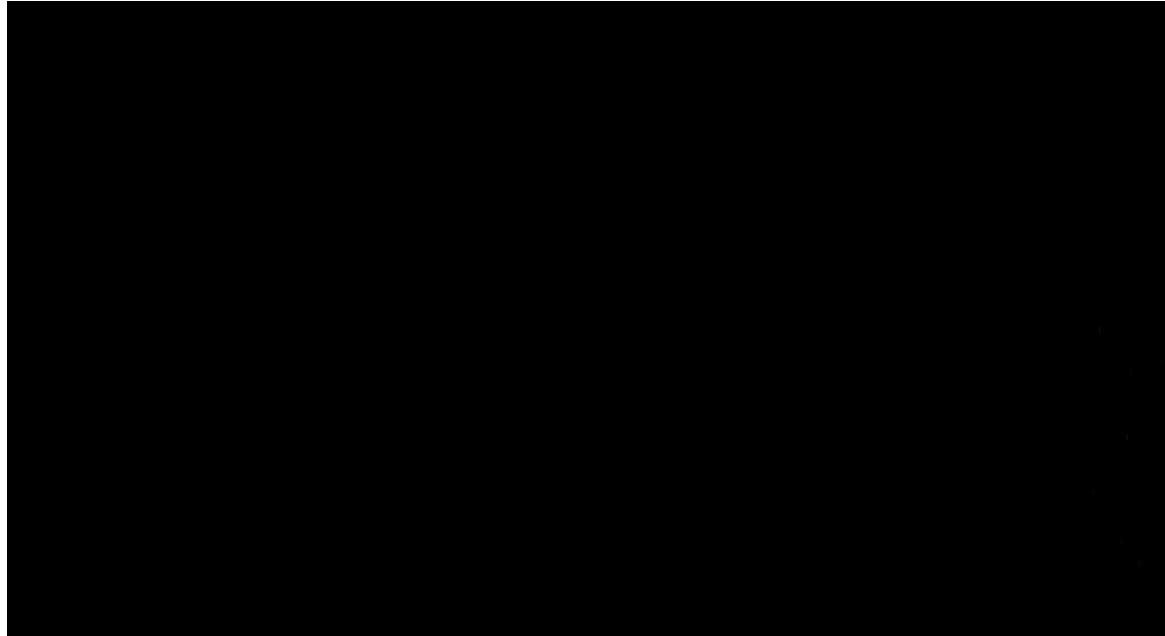


- Multi-modal: [VLEU](#), [OSPC](#)
- Multi-lingual: [KddRES](#), [Self-Denote](#)
- Explainable AI: [ReadPrompt](#), [K-Dial](#)
- .....



# Multi-lingual

## First Cantonese-based Task-oriented DS

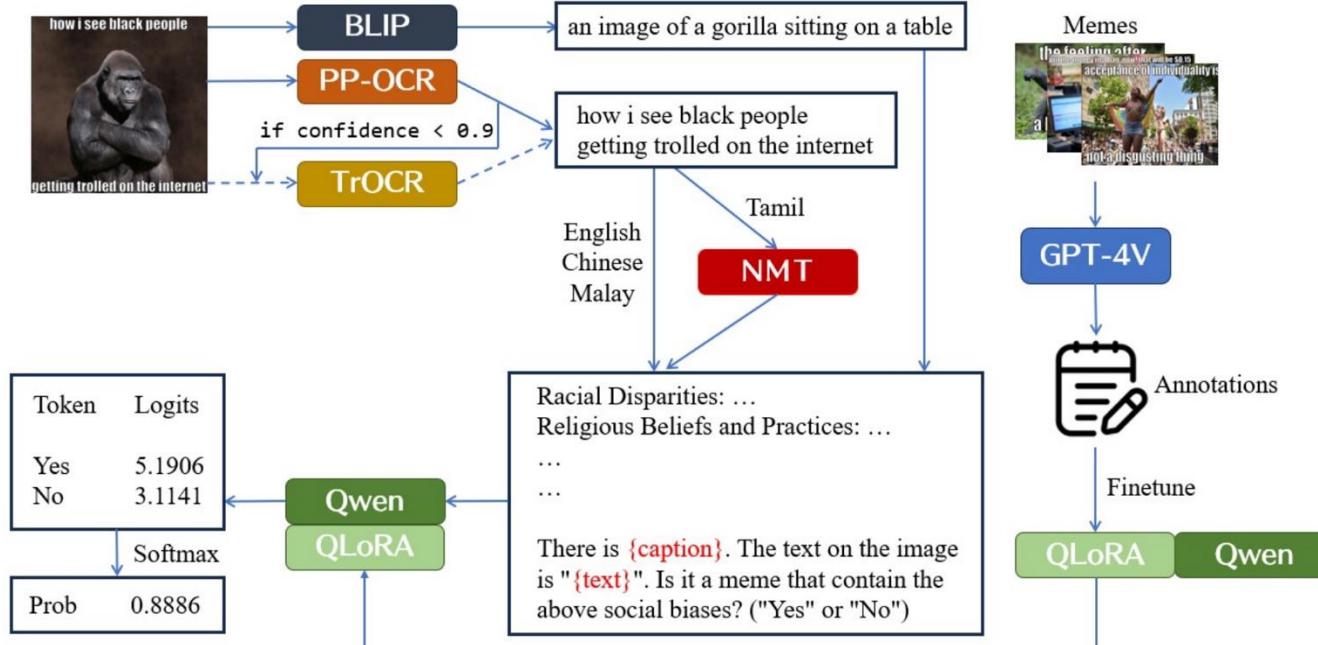


Type	Multi-domain		Single-domain				
	CrossWOZ	MultiWOZ	DSTC2	KVRET	WOZ2.0	CamRest676	KddRES
Language	Chinese	English	English	English	English	English	Cantonese
Speakers	H2H	H2H	H2M	H2H	H2H	H2H	H2H
Dialogues	5012	8438	1612	2425	600	676	834
Turns	84692	115424	23354	12732	4472	1500	8022
Avg.turns	16.9	13.7	14.5	5.3	7.5	2.2	9.5
Slots	72 (10)	25 (12)	8	13	4	7	26

We propose **the first Cantonese fully annotated human-to-human task-oriented dialogue dataset** for restaurants — KddRES. It contains 834 multi-turn dialogues with approximately 8000 utterances and 26 distinct slots with a hierarchy of coarse- to fine-grained entity types.



# Multi-modal / Multi-lingual Champion Solution @OSPC WWW2024



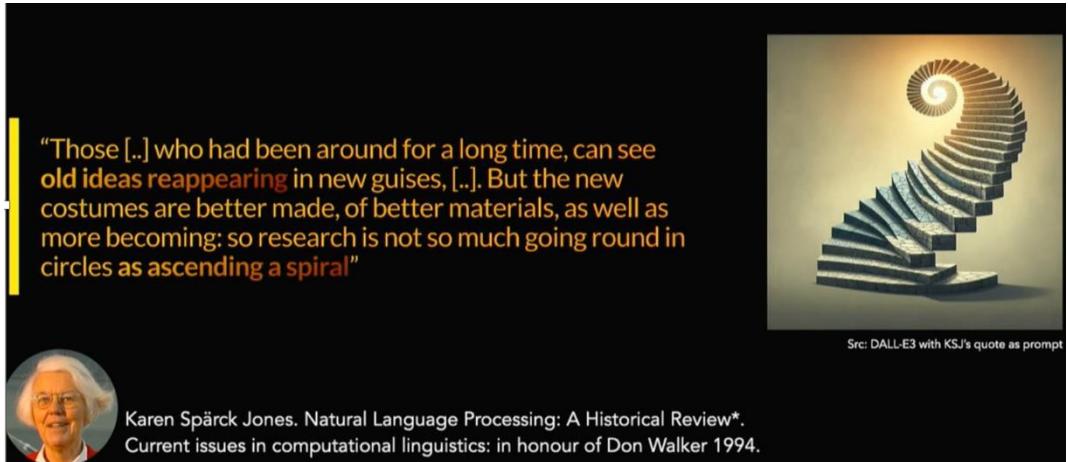
Public Leaderboard

Ranks	Team Name + Total Submissions	AUROC	Accuracy
1	<b>Alris</b> 13 Submission(s) Best public score submission on: 2024-04-11, 09:35:57 (GMT+08:00)	0.7749	0.7087
2	<b>goldbach</b> 15 Submission(s) Best public score submission on: 2024-04-06, 10:38:34 (GMT+08:00)	0.7600	0.5618
3	<b>Baseline</b> 13 Submission(s) Best public score submission on: 2024-04-08, 17:42:42 (GMT+08:00)	0.7597	0.6913

- Effectively detect harmful memes in Singapore is challenging, considering diversity of **social contexts, languages, colloquialisms, and experiences**.
- We propose powerful multimodal, multi-lingual, zero-shot solution based on Qwen, achieves top-1 performance.

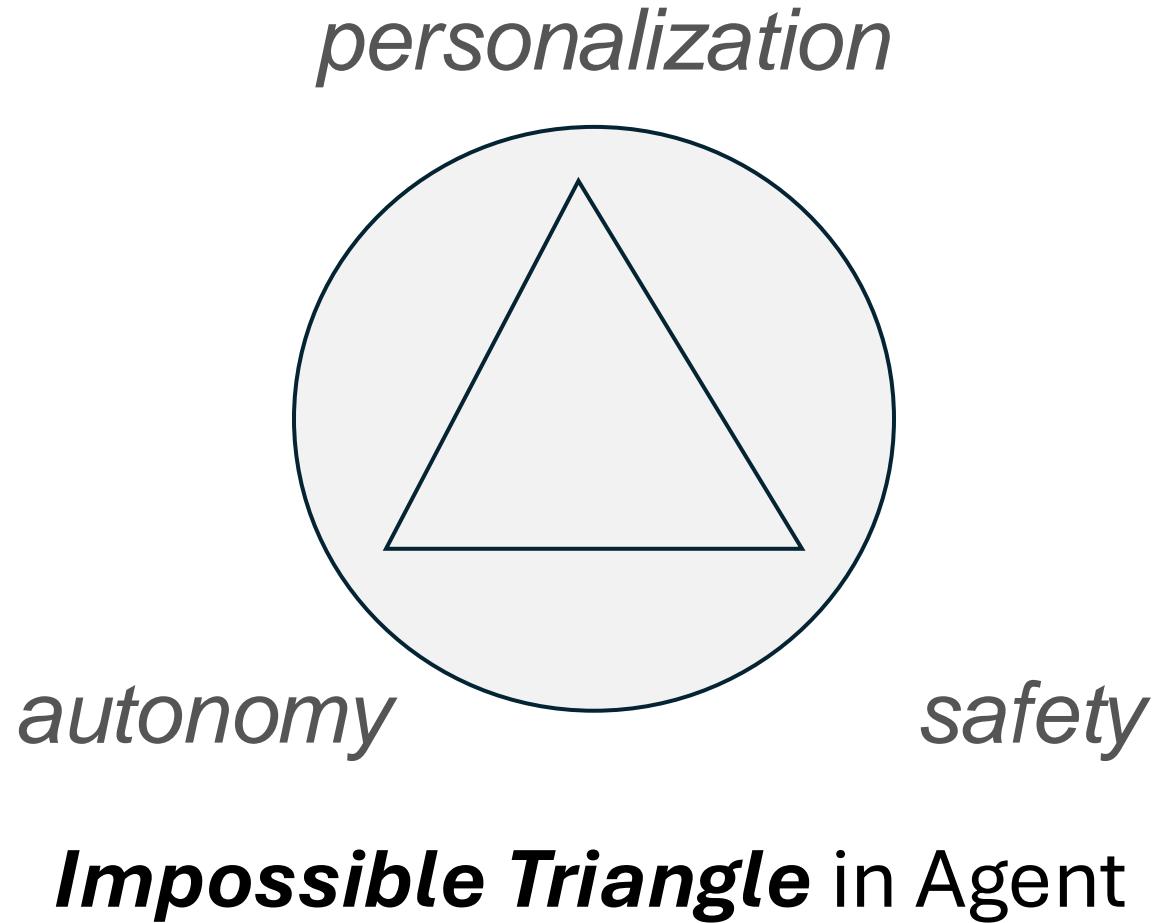


# Future Direction



- Task-oriented Dialogue System (DS)
  - NLU / DST / DPL / NLG
- Open-domain Dialogue System (DS)
  - RAG / Search
- Unified DS / LLM-based DS
- Tool Learning
- Language Agent (*new* Task-oriented DS)
  - Such as Tau-bench / AppBench
- .....

Time



# How my work fit Information Hub

HUBS AND THRUST AREAS

## INFORMATION HUB

The primary goal of the Information Hub is to promote excellence in the fields of information science and technology by facilitating education and cutting-edge research to further the economic and social development of the country. The hub consists of four key thrusts, namely Artificial Intelligence (AI), Data Science and Analytics (DSA), Internet of Things (IoT), and Computational Media and Arts (CMA), which work collaboratively together in a streaming pipeline to develop students' skills and drive innovative research.

INFORMATION HUB

## Vision

To address global challenges arising from human interactions with information and technology in today's era of digital transformation.

INFORMATION HUB

## Mission

To provide world-class education and conduct cutting-edge research with practical applications in information science and technology that not only advance regional development, but also make a global impact.

- **Research:** Dialogue is fundamental interactions of human-human and human-AI, and dialogue system is key research direction at NLP, such as ChatGPT. Moreover, my research on tool learning and language agents aligns closely with vision to address the complexities of *human-AI and AI-AI* interactions.
- **Applications:** My research mostly problem-driven and application-driven. It can be applied into lots of commercial products such as chatbots.
- **Interdisciplinary Collaboration:** Inspired by lots of works from cognitive science, embodied AI, reinforcement learning and other areas, my research aligns well with the *mission* of fostering intelligent, impactful systems by addressing real-world challenges in both digital and physical contexts.



# Acknowledgements

## In Group

Kam-Fai Wong, Professor, CUHK  
Wenjie (Maggie) Li, Professor, PolyU  
Rufeng Xu, Professor, HIT (SZ)  
Zhongyu Wei, Associate Professor, Fudan  
Lingzhi Wang, Associate Professor, HIT (SZ)  
Jing Li, Assistant Professor, PolyU  
Jing Ma, Assistant Professor, HKBU  
.....

Also lots of students / mentors:

Huimin Wang, Zhiming Mao, Jingtao Cao, Zezhong Wang,  
Boyang Xue, Rui Wang, Liang Chen, Yiming Du, Wai-  
Chung Kwan, ...

## Outside of Group

Jeff Z. Pan, Professor, University of Edinburgh  
Irwin King, Professor, CUHK  
Qun Liu, Professor, Huawei Noah's Ark  
Yue Zhang, Professor, Westlake University  
Mirella Lapata , Professor, University of Edinburgh  
Tat-Seng Chua, Professor, NUS  
Yefeng Zheng, Professor, Westlake University  
Yankai Lin, Associate Professor, Renmin University of China  
Yang Deng, Assistant Professor, SMU  
Zeming Liu, Assistant Professor, Beihang University  
Pasquale Minervini, Assistant Professor, University of Edinburgh  
Yu Li, Assistant Professor, CUHK  
...

Also lots of students / mentors:

Fei Mi, Minda Hu, Wenyu Huang, Yu Zhao, Baohang Zhou,  
Rongwu Xu, Jianqiao Lu, Cunxiang Wang, Heming Xia, ...



# Publications

- **Internal Cognitive Tool (Conversational Strategies, Reasoning Modules, Meta-reasoning, ...)**
  - Cue-CoT: Chain-of-thought Prompting for Responding to In-depth Dialogue Questions with LLMs (**EMNLP 2023**)
  - Enhancing Large Language Models Against Inductive Instructions with Dual-critique Prompting (**NAACL 2024**)
  - Role Prompting Guided Domain Adaptation with General Capability Preserve for Large Language Models (**NAACL 2024**)
  - PerLTQA: A Personal Long-Term Memory Dataset for Memory Classification, Retrieval, and Fusion (**Best Paper @ SIGHAN of ACL 2024**)
  - AutoPSV: Automated Process-Supervised Verifier (**NeurIPS 2024**)
- **External Physical Tool (Search Engine, Knowledge Sources, Actions, ...)**
  - Integrating Pretrained Language Model for Dialogue Policy Learning (**ICASSP 2022**)
  - Large Language Models as Source Planner for Personalized Knowledge-grounded Dialogues (**EMNLP 2023 Best Paper @ International Doctoral Forum**)
  - UniRetriever: Multi-task Candidates Selection for Various Context-Adaptive Conversational Retrieval (**LREC-COLING 2024**)
  - Empowering Large Language Models: Tool Learning for Real-World Interaction (**ToolsMeetLLM @SIGIR 2024**)
- **Framework (Planning, Tool Calls, Tool Conflicts, ...)**
  - AppBench: Planning of Multiple APIs from Various APPs for Complex User Instruction (**EMNLP 2024**)
  - Knowledge Conflicts for LLMs: A Survey (**EMNLP 2024**)
  - Analysing the Residual Stream of Language Models Under Knowledge Conflicts (**MINT @NeurIPS 2024**)
  - ...



# Thank You!

(Q & A)

