



How is Today's Weather? Human-AI Interaction Makes the Instruction Clear

ESTR 4999 Presentation

Supervisor: Professor Michael Lyu

Presenter: SHI Juluan (1155160208)

Chan Chun Yip (1155158514)

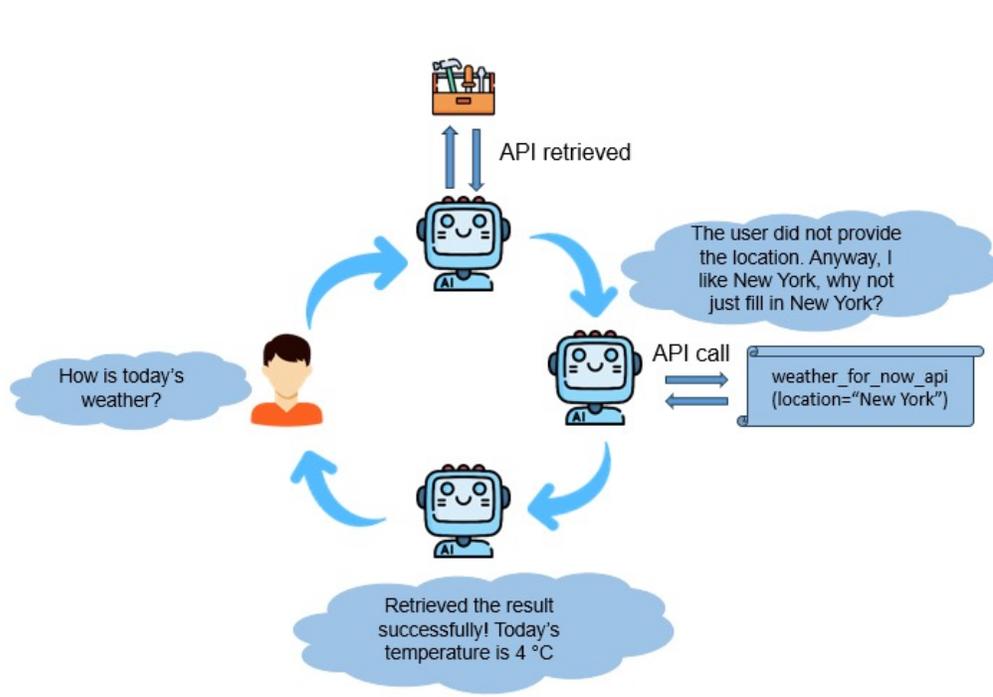


香港中文大學

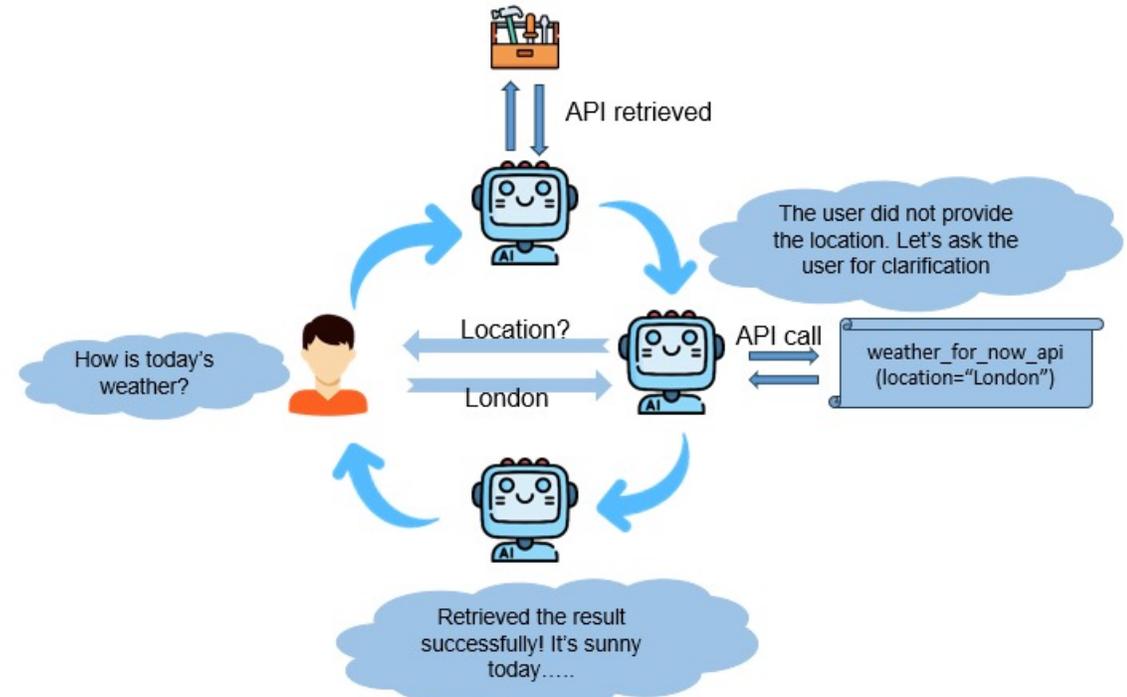
The Chinese University of Hong Kong



Motivation



Original

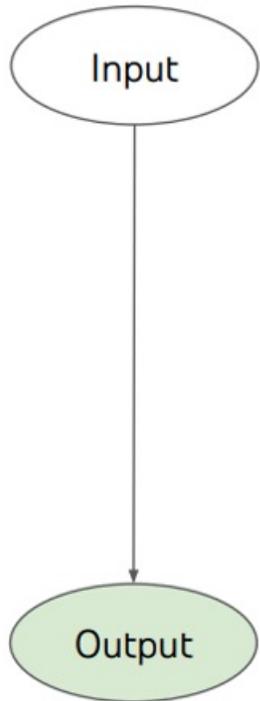


Expectation

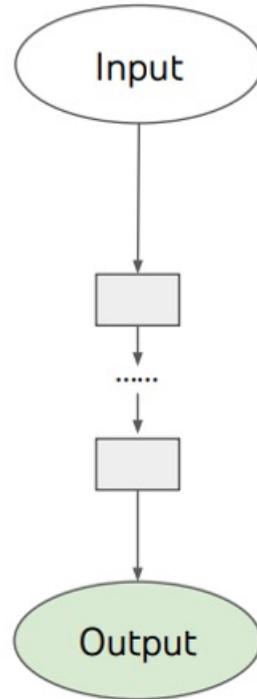


➔ Background

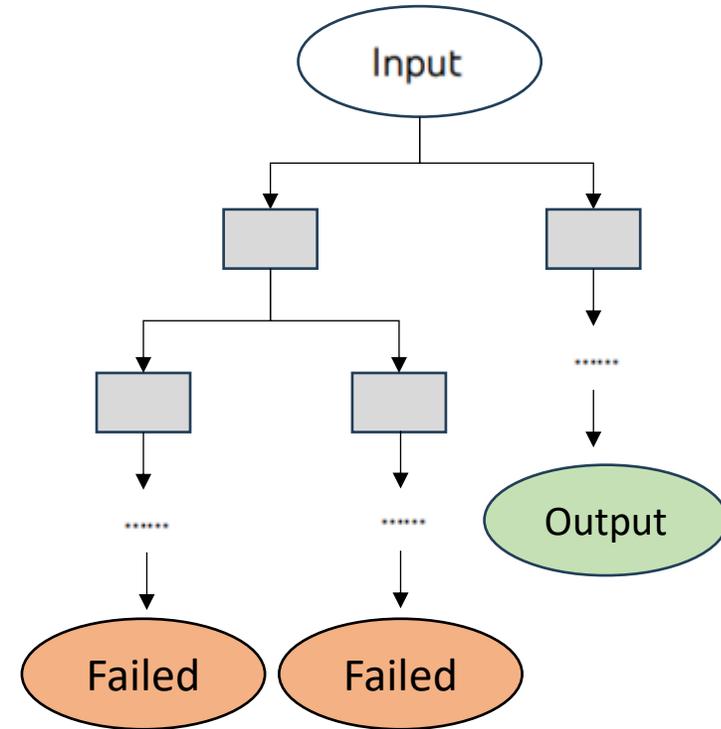
Prompting for decision-making



(a) Input-Output Prompting (IO)



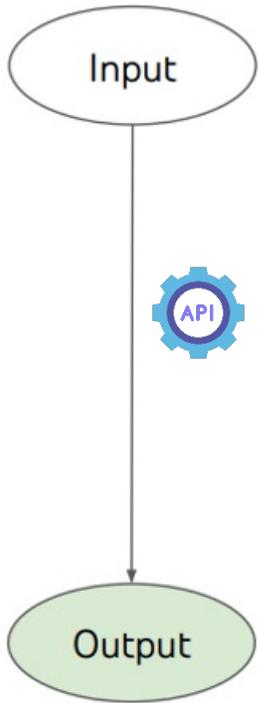
(b) Chain of Thought Prompting (CoT)



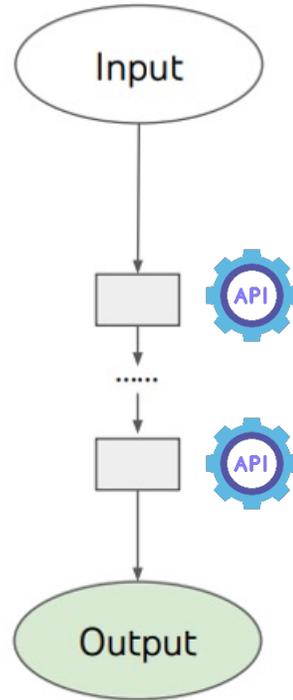
(c) Depth-First Search-based Decision Tree (DFSDT)



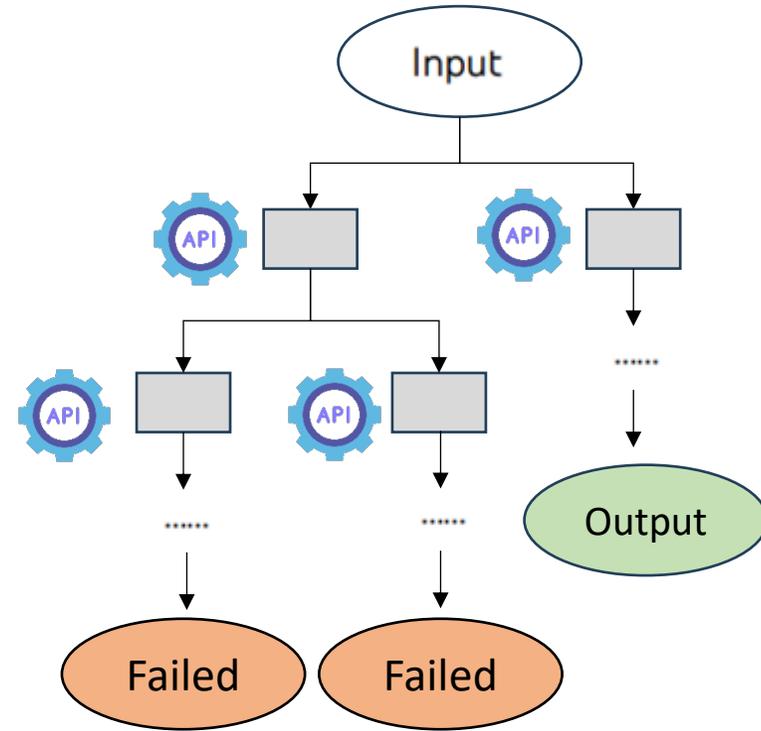
➔ Tool Learning



(a) Input-Output Prompting (IO)



(b) Chain of Thought Prompting (CoT)



(c) Depth-First Search-based Decision Tree (DFSDT)



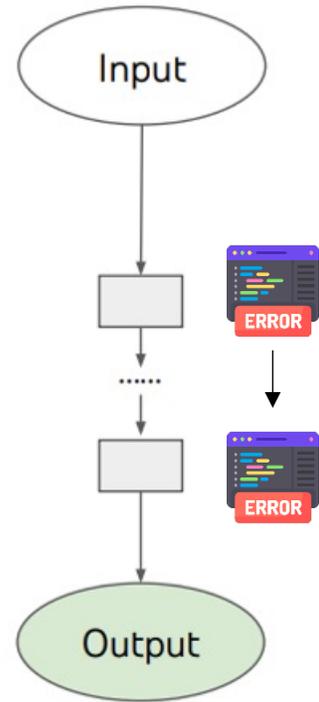
➔ Special Feature of APIs



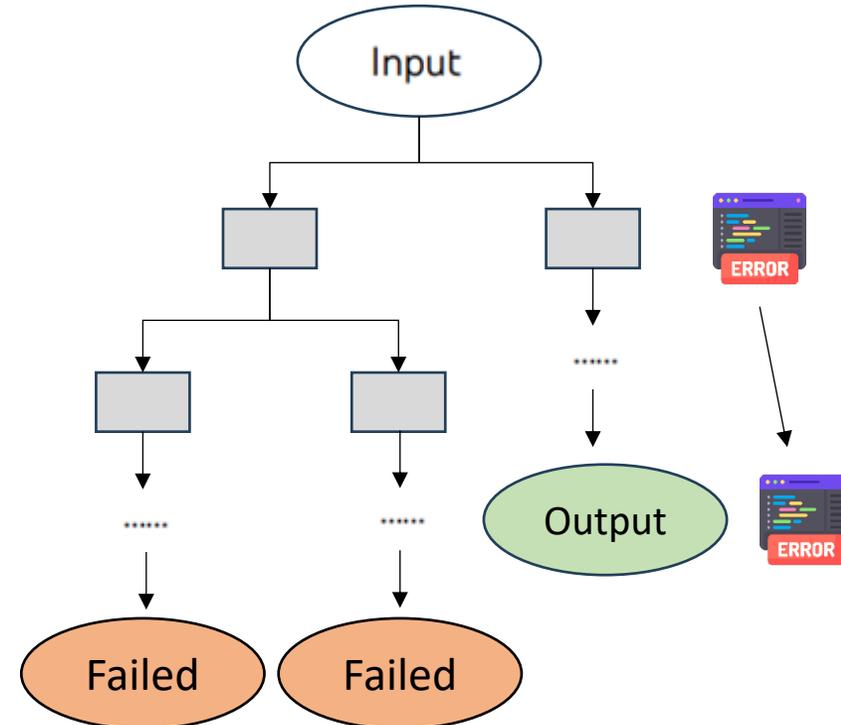
```
“Category name”: “Finance”  
“Tool name”: “Exchange Rate”  
“API name”: “historical”  
“API description”: “get historical rates for any day since 1999”  
“Required parameters”:[  
  {  
    “name”: “date”  
    “type”: “string”  
    “default”: “2000-01-03”  
  }  
]
```



➔ Error Propagation



(b) Chain of Thought Prompting (CoT)



(c) Depth-First Search-based Decision Tree (DFSDT)



Contributions

- Conduct an extensive examination of the failures in tool utilization by current LLMs when faced with ambiguous user instructions in **real-life scenarios**, categorizing the prevalent issues into **four** distinct categories.
- Crafted our benchmark, which thoughtfully incorporates the above **four** distinct categories of challenges alongside a **diverse** selection of tools.
- Refined the existing planning and reasoning algorithms and introduced a **novel algorithm**, termed **QwN**. This algorithm is designed to prompt LLMs to **actively request clarifications** from users upon facing **uncertainties**.
- Devised innovative **evaluation metrics** tailored to the new tasks and introduced an **automated evaluation method** to streamline the efficient assessment of results.



➤ Categories of User Instruction Errors

- Classify these instructions into 4 categories based on their characteristics.
- **Instructions missing key details (IKEI)**
 - User instructions that omit crucial details



How is today's weather

- **Instructions with unclear references (IUR)**

- User instructions include elements that can be interpreted in several ways



I want to know the director of the movie 'The Matrix'



➤ Categories of User Instruction Errors

- **Instructions with errors (IWE)**

- Information is incorrect



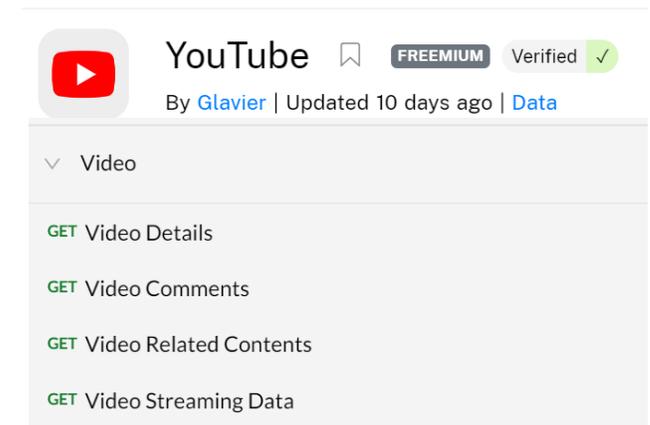
Please help me to log in to my Twitter. My user account is 'abcde@gmail.com' and the password is '123456'

- **Instructions beyond tool capabilities (IBTC)**

- Request actions or answers beyond the available APIs



Please post a comments “...”





➤ User Instruction Analysis

Overview of API Categories Analyzed

- Number of Categories: 28
- Total API Sets Reviewed: Over 150



Type of error	Error percentage
Information missing	56%
Information unclear	11.3%
Information incorrect	17.3%
Tool limitation	15.3%

Table 1: Error percentages in various problematic instructions.

Many APIs work as database

API dojo.net

Weather

🔖

FREEMIUM

By [Api Dojo](#) | Updated 10 days ago | [Weather](#)

▼ Required Parameters

date	20200622
STRING	REQUIRED Date to get weather forecast, format as yyyyMMdd
latitude	37.765
NUMBER	REQUIRED The latitude of GEO location to get weather forecast
longitude	-122.463
NUMBER	REQUIRED The longitude of GEO location to get weather forecast



➔ Crafted Benchmark

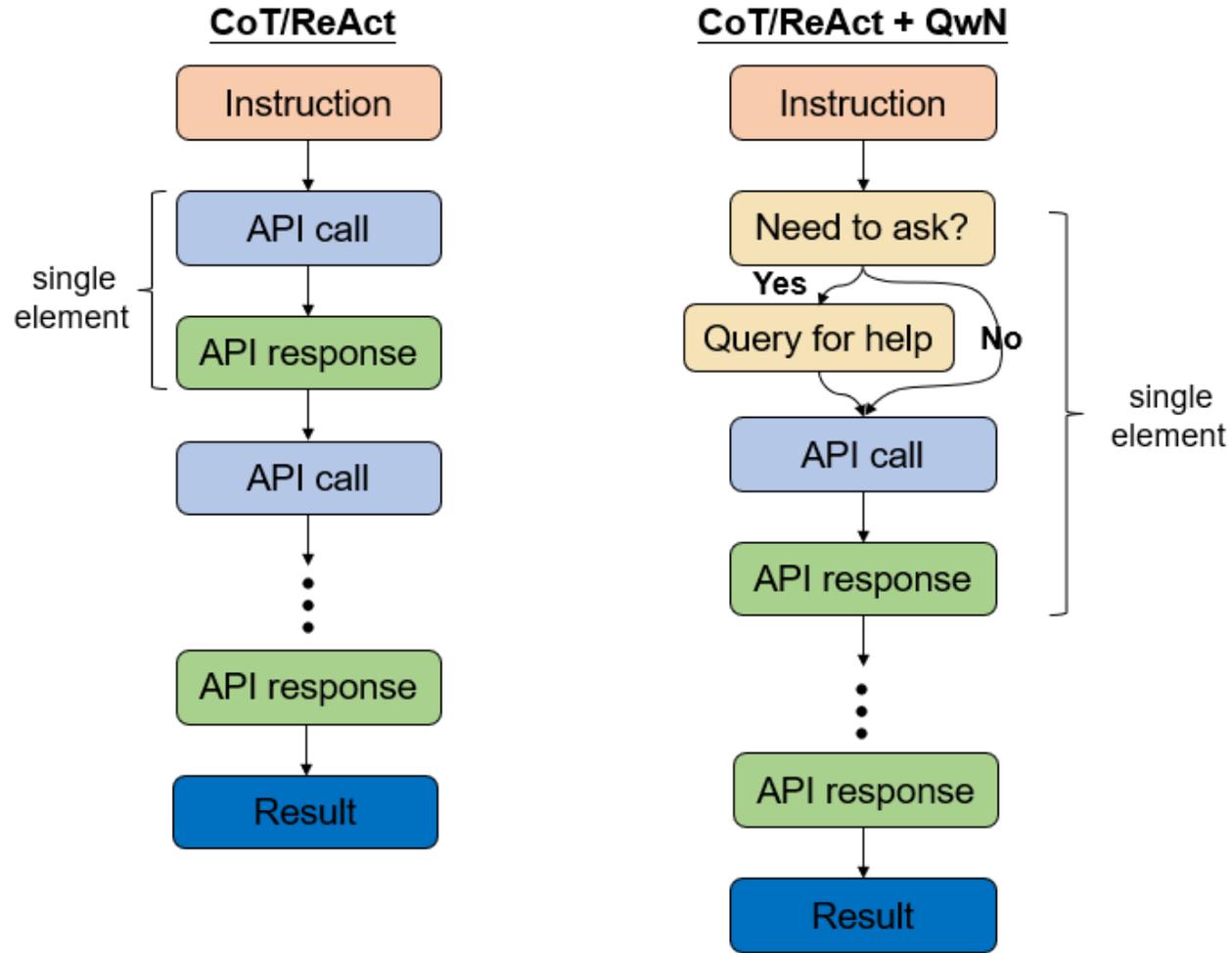
- Curated our benchmark around these four categories, with each category comprising 50 user instructions.

```
"api_list": [  
  {  
    "category_name": "Weather",  
    "tool_name": "weather forecast 14 days",  
    "api_name": "Get forecastdata by lat/lon",  
    "api_description": "get forecast for 14 days for the location Lat/Lon",  
    "required_parameters": [  
      {  
        "name": "LAT",  
        "type": "NUMBER",  
        "description": "Latitude",  
        "default": "51.5"  
      }  
    ]  
  }  
]
```

```
"query": "I want to know the weather forecast for the next 14 days. Can you provide me with the temperature, dew point, relative humidity, pressure, wind speed and direction, rainfall, and symbol for each day?",  
"question need to be asked": "Please specify the city you want the weather forecast",  
"clarification": "London",  
"expected API calling": [  
  {  
    "name": "search_location_by_name_or_zip_code_for_weather_forecast_14_days",  
    "arguments": "{\n  \"city\": \"London\"\n}"  
  }  
],  
"relevant APIs": [  
  "weather forecast 14 days",  
  "Search location by Name or zip code"  
]
```



➔ Query-when-Need Prompting





➔ Evaluaiton Metrics

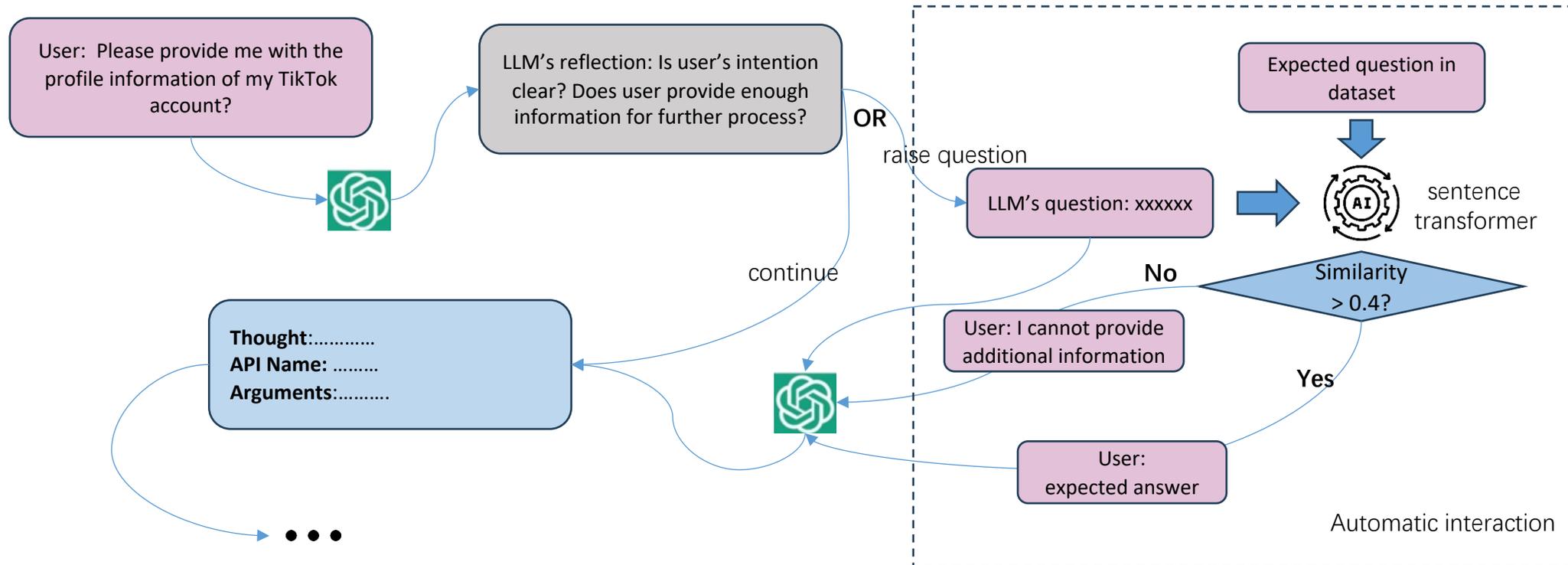
- **Goal 1:** This metric evaluates the capability of LLMs to ask the anticipated questions that pinpoint the ambiguous elements in user instructions.
- **Goal 2:** This metric assesses the ability of LLMs to use all available information to invoke the correct API calls.
- **Goal 3:** This metric measures the ability of LLMs to extract the anticipated information from previous API calls to fulfil the user's instructions.
- **Average Redundant Asked questions (Re):** This metric evaluates the quantity of irrelevant or redundant questions asked by tool-augmented LLMs during the instruction executing process.
- **Steps:** This metric quantifies the average number of actions required to complete an instruction.

Accuracy

Efficiency



➔ Evaluation Pipeline --- Auto Interaction





Evaluation Pipeline --- Auto Evaluation

```

{
  "role": "assistant",
  "content": "Question: Could you please provide me with your friend's LeetCode username?"
},
{
  "role": "user",
  "content": "His LeetCode username is 'michaelj103'"
},
{
  "role": "assistant",
  "content": null,
  "function_call": {
    "name": "getmediumquestion_for_leetcodeapi",
    "arguments": "{\n  \"username\": \"michaelj103\"\n}"
  }
},
{
  "role": "function",
  "name": "getmediumquestion_for_leetcodeapi",
  "content": "{\n  \"error\": \"\", \"response\": \"515\"\n}"
},
{
  "role": "assistant",

```

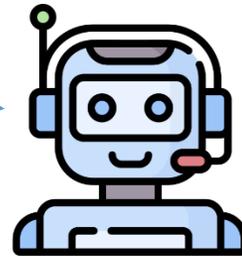
history of instruction execution

```

"expected API calling": [
  {
    "name": "getmediumquestion_for_leetcodeapi",
    "arguments": "{\n  \"username\": \"michaelj103\"\n  }"
  },
  {
    "name": "geteasyquestion_for_leetcodeapi",
    "arguments": "{\n  \"username\": \"michaelj103\"\n  }"
  },
  {
    "name": "gethardquestion_for_leetcodeapi",
    "arguments": "{\n  \"username\": \"michaelj103\"\n  }"
  }
],

```

expected API calls



Auto Evaluator (ChatGPT4)

query_id	query	steps	redundant	goal1	goal2	goal3	reason
1	As a frequent player o	3	0	Success	Success	Success	Reason: For Objective 2, the model successfully invoked t correct arguments. The actual function call made arguments '{"date1":"12-01-2023","date2":"12-01-2023"}'
2	I want to know my Le	5	0	Success	Success	Success	Reason: For Objective 2, the model successfully invoked i
3	I want to know the Le	5	0	Success	Success	Success	Reason: For Objective 2, the model successfully invoked i
4	I want to start my day	4	0	Success	Success	Failure	Reason: For Objective 2, the model correctly invoked the 'getmultiplequestions_for_asktheworld'. The rec
5	I need to conduct a re	5	1	Success	Failure	Failure	Reason: For Objective 2, the model invoked two function 'getmultiplequestions_for_asktheworld'. The rec model, but the arguments are not the same. The Reason: For Objective 2, the model invoked the function

evaluation result



Main Experiments

Table 2: Assessing the accuracy of various LLMs using different prompting methods in our benchmark

base model	framework	IKEI			IUR			IWE			IBTC		
		G1(%)	G2(%)	G3(%)									
gpt-3.5-turbo	CoT	0.34	-	-	0.24	-	-	0.21	-	-	-	-	-
	+ QwN	0.86	0.56	0.48	0.76	0.42	0.24	0.58	0.28	0.16	-	-	-
-16k-0613	DFS	0.53	-	-	0.42	-	-	0.42	-	-	-	-	-
	+ QwN	0.84	0.44	0.52	0.88	0.46	0.42	0.86	0.26	0.4	-	-	-
gpt-4-0125-preview	CoT	0.68	-	-	0.58	-	-	0.35	-	-	-	-	-
	+ QwN	0.84	0.50	0.44	0.76	0.42	0.34	0.42	0.26	0.30	-	-	-
	DFS	0.64	-	-	0.56	-	-	0.37	-	-	-	-	-
	+ QwN	0.8	0.46	0.70	0.64	0.42	0.36	0.51	0.40	0.58	-	-	-

Accuracy

Table 3: Assessing the efficiency of various LLMs using different prompting methods in our benchmark.

base model	framework	IKEI		IUR		IWE		IBTC	
		Re	Steps	Re	Steps	Re	Steps	Re	Steps
gpt-3.5-turbo	CoT	-	4.64	-	4.52	-	5.09	-	3.10
	+ QwN	0.84	5.80	1.46	6.16	1.67	6.28	-	3.75
-16k-0613	DFS	-	50.20	-	55.50	-	45.83	-	17.06
	+ QwN	10.90	68.60	11.54	74.50	15.30	79.80	-	9.25
gpt-4-0125-preview	CoT	-	3.00	-	2.92	-	2.79	-	1.58
	+ QwN	0.18	3.70	0.20	3.40	0.23	3.47	-	1.17
	DFS	-	5.26	-	6.88	-	6.18	-	2.65
	+ QwN	0.34	5.30	0.36	4.98	0.37	6.76	-	1.25

Efficiency



➔ Main Result (Accuracy)

base model	framework	IKEI			IUR			IWE			IBTC		
		G1(%)	G2(%)	G3(%)									
gpt-3.5-turbo	CoT	0.34	-	-	0.24	-	-	0.21	-	-	-	-	-
	+ QwN	0.86	0.56	0.48	0.76	0.42	0.24	0.58	0.28	0.16	-	-	-
-16k-0613	DFS	0.53	-	-	0.42	-	-	0.42	-	-	-	-	-
	+ QwN	0.84	0.44	0.52	0.88	0.46	0.42	0.86	0.26	0.4	-	-	-
gpt-4-0125-preview	CoT	0.68	-	-	0.58	-	-	0.35	-	-	-	-	-
	+ QwN	0.84	0.50	0.44	0.76	0.42	0.34	0.42	0.26	0.30	-	-	-
	DFS	0.64	-	-	0.56	-	-	0.37	-	-	-	-	-
	+ QwN	0.8	0.46	0.70	0.64	0.42	0.36	0.51	0.40	0.58	-	-	-

- **Interaction module encourages LLM to interact with users to solve ambiguous instruction**

-This can be particularly seen in the G1 scores across both models when the interaction module is utilized

- **Evaluating IBTC differently**

-We do not conduct statistical analysis for groups G1, G2, G3, and Re. Instead, our focus shifts to comparing the Steps



➔ Main Result (Efficiency)

base model	framework	IKEI		IUR		IWE		IBTC	
		Re	Steps	Re	Steps	Re	Steps	Re	Steps
gpt-3.5-turbo	CoT	-	4.64	-	4.52	-	5.09	-	3.10
	+ QwN	0.84	5.80	1.46	6.16	1.67	6.28	-	3.75
-16k-0613	DFS	-	50.20	-	55.50	-	45.83	-	17.06
	+ QwN	10.90	68.60	11.54	74.50	15.30	79.80	-	9.25
gpt-4-0125-preview	CoT	-	3.00	-	2.92	-	2.79	-	1.58
	+ QwN	0.18	3.70	0.20	3.40	0.23	3.47	-	1.17
	DFS	-	5.26	-	6.88	-	6.18	-	2.65
	+ QwN	0.34	5.30	0.36	4.98	0.37	6.76	-	1.25

- **Impact of the interaction module on the average number of steps**
 - In every instance of adding an interaction module to the base model framework, there is an increase in the number of steps required to complete an instruction
- **DFS method introduces more redundant questions**
 - DFS method on gpt-3.5 introduces a significantly higher amount of redundancy
- **GPT 4 avoid asking unnecessary questions**
 - GPT-4 is effective in identifying and inquiring about unclear elements without generating excessive unnecessary questions, as indicated by the lower "Re "



➔ Case study

- **Instructions with errors (IWE)**
- **Instructions beyond tool capabilities (IBTC)**

Instructions with errors

User: I'm a blogger and I want to write a detailed review of a specific product with SKU X005789621. Can you provide me with all the information about this product and provide reviews for these products?

Assistant : Action call: "products_detail_for_wayfair" {sku: X005789621}
{error: request invalid, data error. status_code=500, response: }

Assistant : Question: I apologize, but it seems that the request for product information with SKU X005789621 is currently unavailable. Is there another product you would like information about?"

User: "The correct one is 'W003247135'"

Instructions beyond tool capabilities

User: Get subtitles for a Vimeo video with the ID '12345abc

Assistant : Due to the limitation of toolset, I cannot solve the question

Assistant : End



Conclusion

- **Introduced the Interaction-for-Tool-Usage benchmark**
 - Designed to evaluate an LLM's ability to request clarifications, thereby improving its task execution capabilities.
- **Developed the Query-when-Need algorithm**
 - Empowers LLMs to actively seek user input whenever there is uncertainty in the instructions provided
- **Created an automated evaluation tool**
 - Reduces the manual efforts required in performance assessment, facilitating a more efficient and consistent evaluation process.



Thanks for listening!



香港中文大學
The Chinese University of Hong Kong