



# Enabling Interactive Conversation During Tool Utilization

ESTR 4998 Presentation

Supervisor: Professor Michael Lyu

Presenter: SHI Juluan (1155160208)

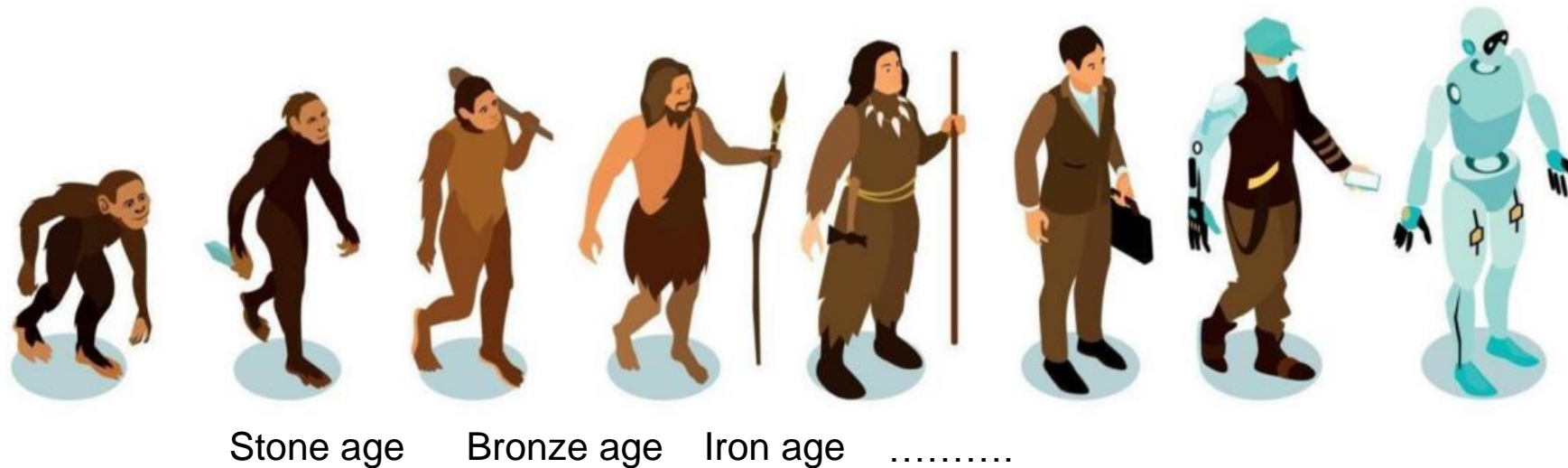
Chan Chun Yip (1155158514)



香港中文大學  
The Chinese University of Hong Kong

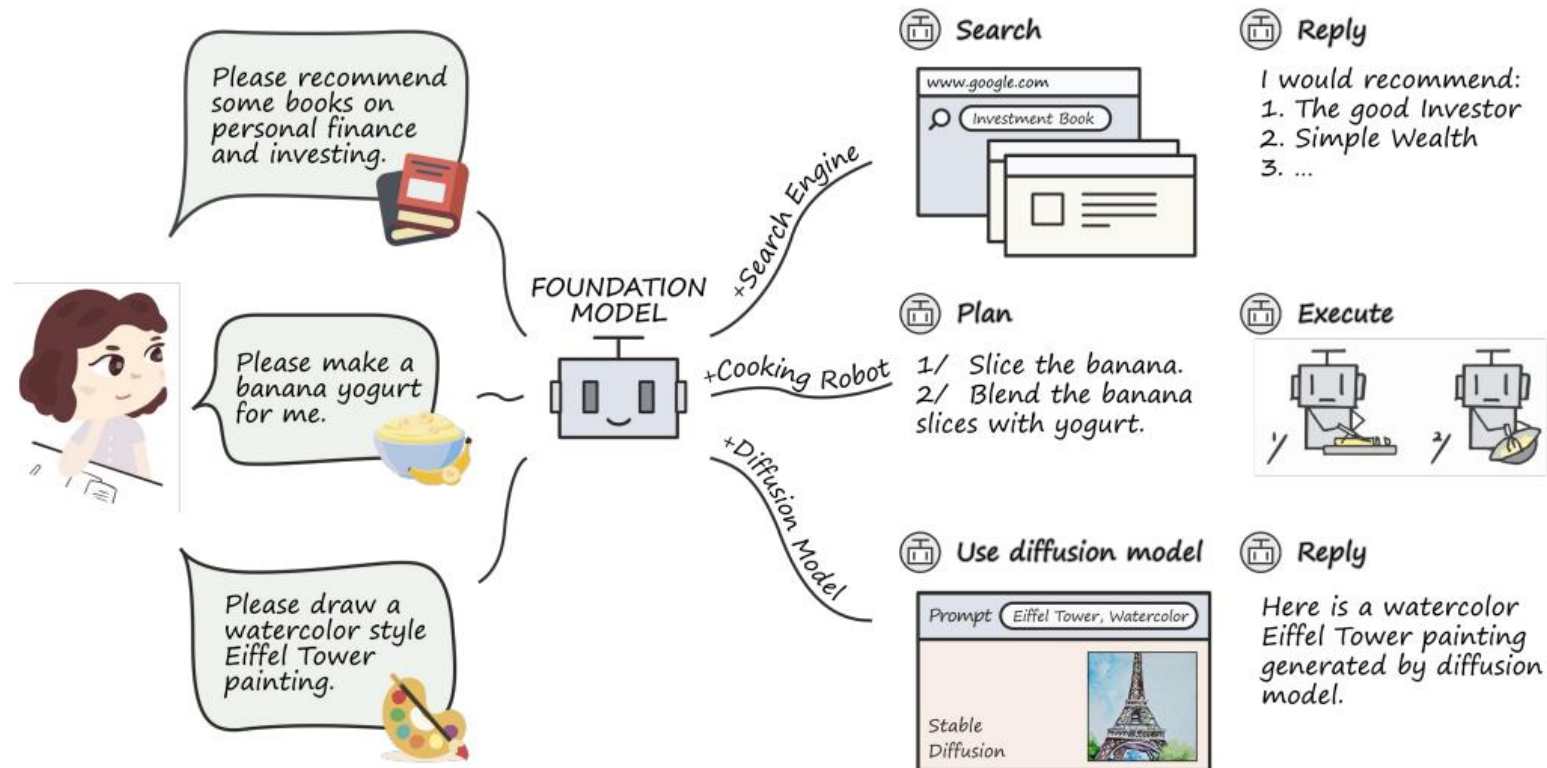
# ➤ Background: Tools and Intelligence

- the creation and usage of tools → humans VS animals
- the usage of tools: a representation of human intelligence
- Challenges: How to enable LLMs the ability to use tools wisely?



# ➤ Background: Tool Learning

- Tool Learning: foundation models can follow human instructions and manipulate tools for task solving





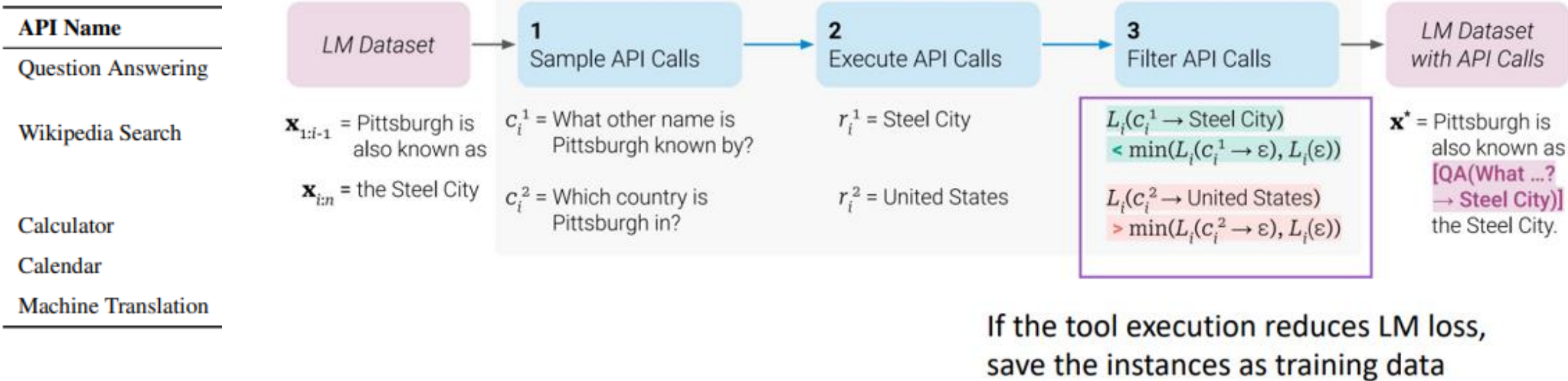


# ➤ Related Work

## Toolformer

### Self-supervised Tool Learning

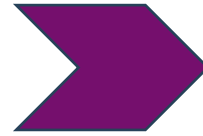
- Pre-defined tool APIs
- Encourage models to call and execute tool APIs
- Design self-supervised loss to see if the tool execution can help language modeling





## ➤ Related Work

### Limitations of Toolformer



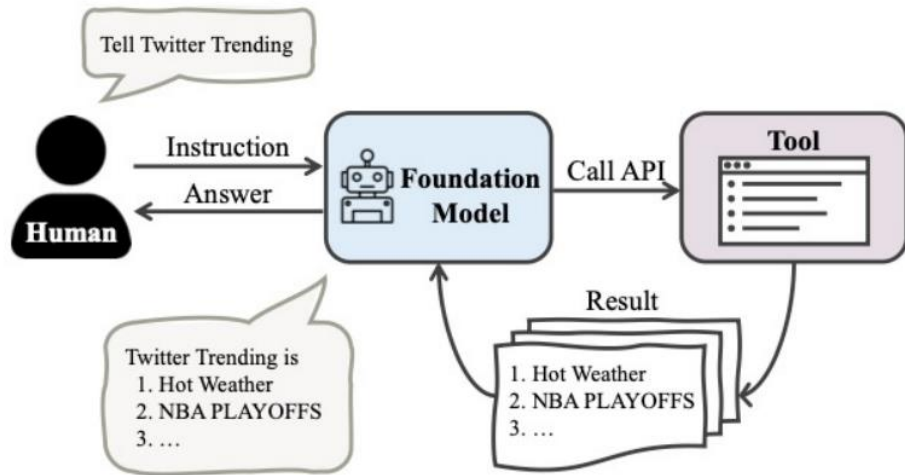
### Requirements of Real-world User Instructions

- Only supports several predefined APIs & cannot scale efficiently
- Only supports the instruction that require one round of API call.

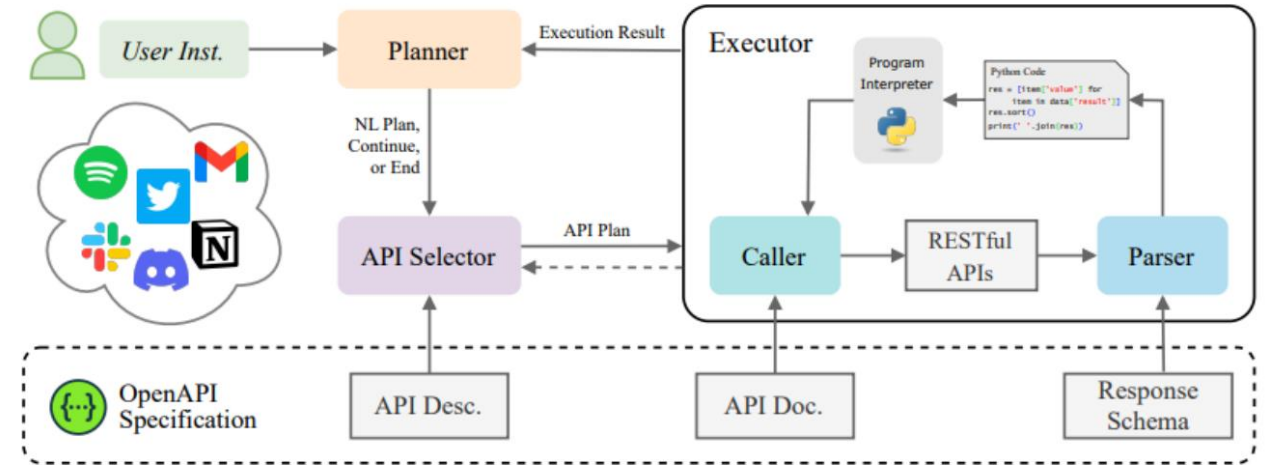
- Planning and reasoning (decompose complex instruction into several solvable subtasks)
- Selecting API from a large-scale toolset base on API description
- Learning to call API correctly after reading API documentation



# Two Recently Proposed Frameworks



ToolLLM (single-agent framework)



RestGPT (multi-agent framework)

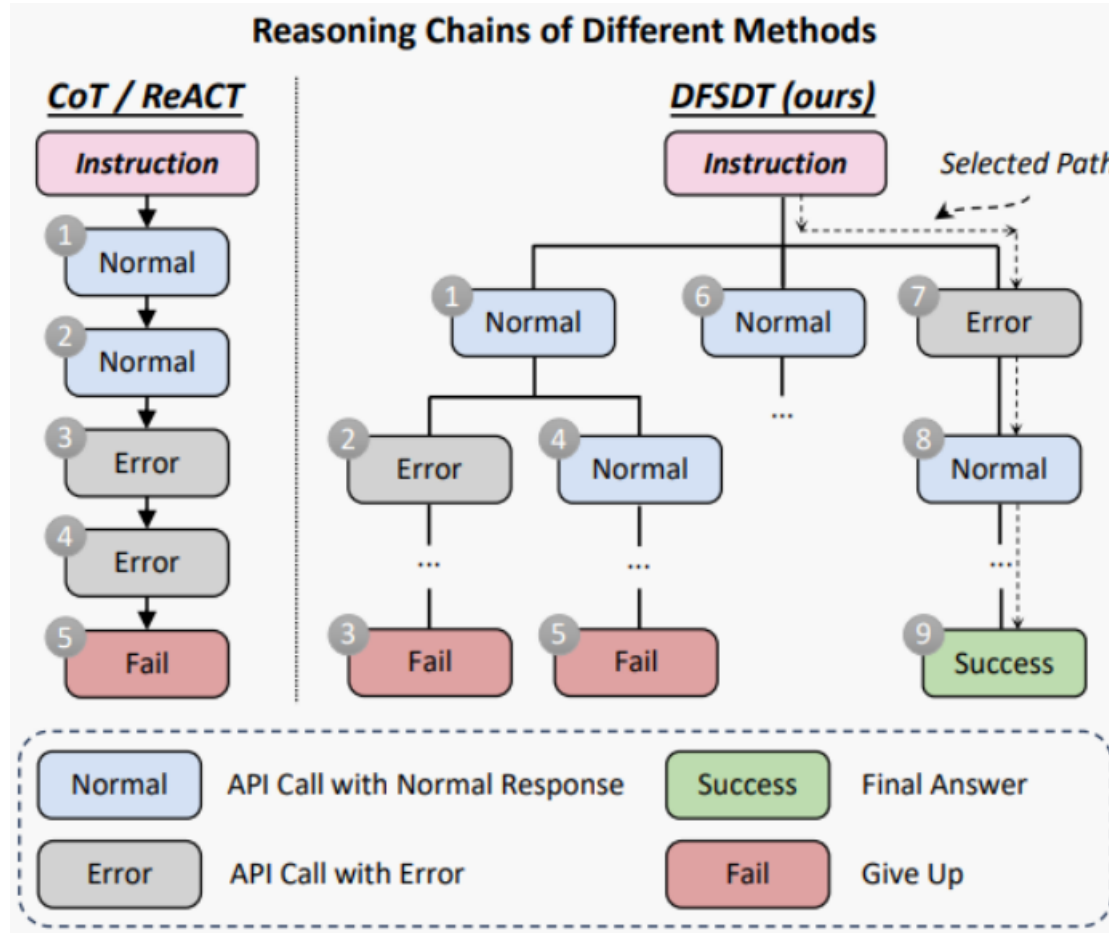
Both frameworks encourage an end-to-end procedure



User Instruction as input & Final Answer as output

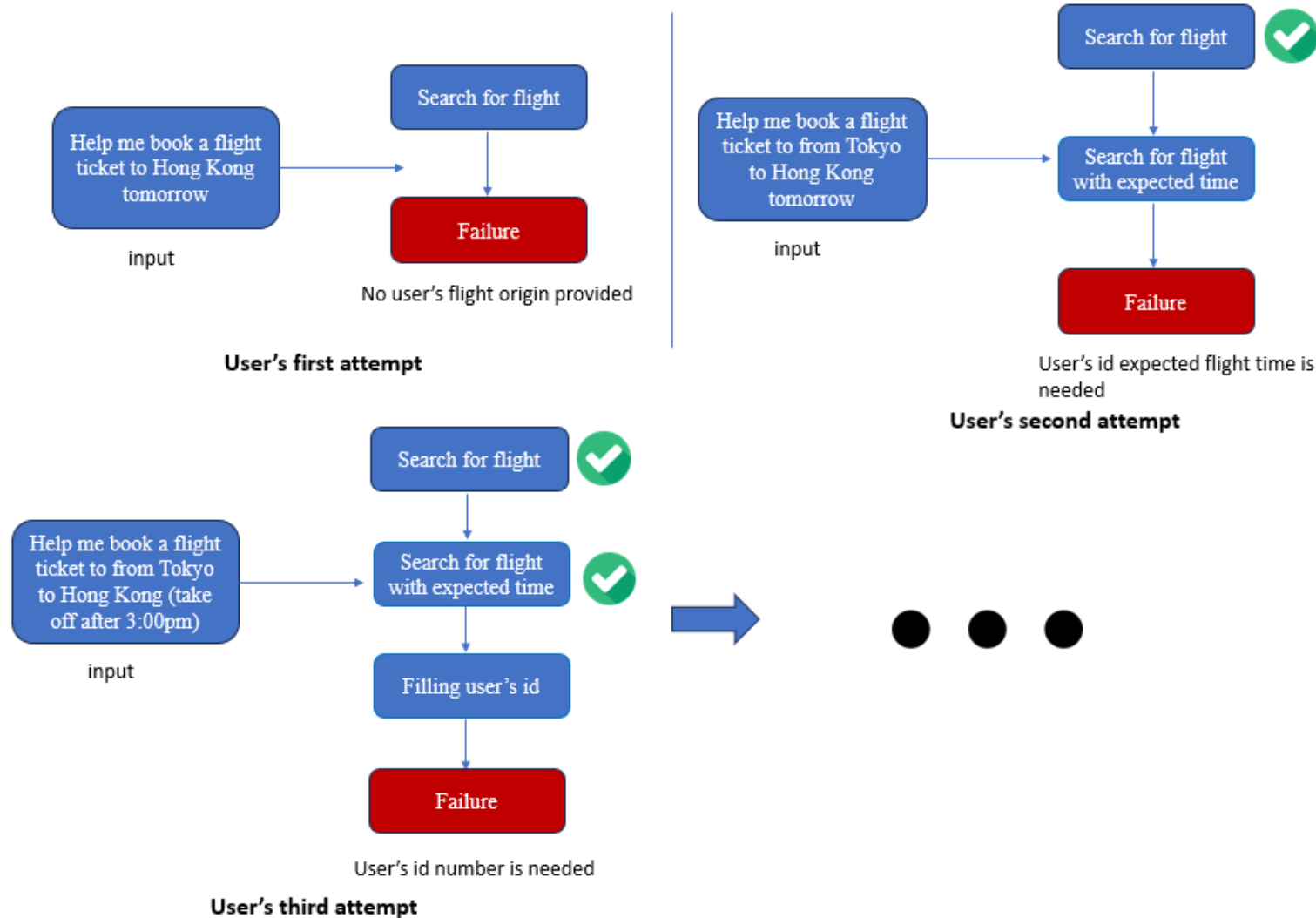


# ➤ Motivation: Limitation of Previous Works



# ➤ Motivation: Limitation of Previous Works

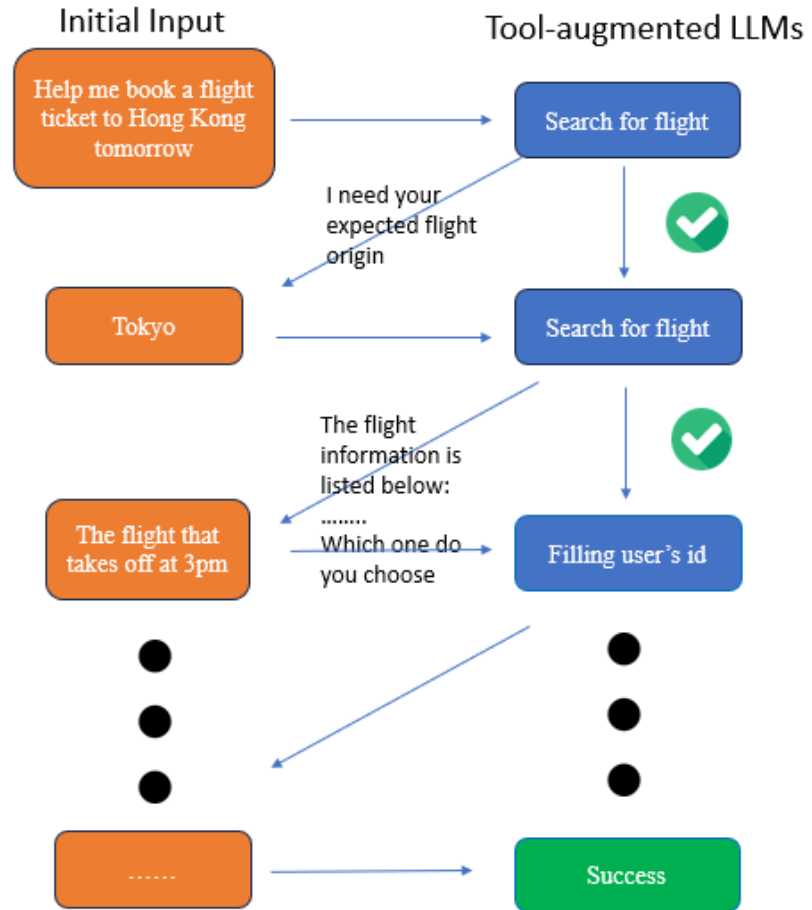
- For most of times, users may not be able to give a complete instruction in a query







# Motivation: Limitation of Previous Works



**Research Target:**  
Enable tool-augmented framework to interact with users during tool utilization



## Our key contributions:

- Conduct an **empirical study** of existing datasets to analyze ambiguous user instructions.
- Curate a **new human-annotated benchmark** to evaluate tool-augmented framework's ability in handling unclear instructions.
- Introduce Query when Need (QwN), **an innovative method** to enable **interaction** between users and agents during the tool invocation process.

# ➤ Problem Analysis of Existing Frameworks: ToolLLaMA

We have investigated ToolLLaMA, a single-agent framework approach:

- 200 failure cases and investigated the solution paths.
- Categorize these failure cases into nine groups.

Type of error	Error percentage	
Information missing	32.5%	54.5 %
Information unclear	8.5%	
Information incorrect	13.5%	
Tool limitation	7.5%	45.4%
API down	28%	
Error in multi-tool	3%	
Code problem	2%	
Wrong tool	4.5%	
Return in complex format	0.5%	

Error analysis of ToolLLaMA

Type of Error	Example
Information Missing	Q: Extract thumbnail images from mp4 videos.  Result: videourl= https://example.com/video.mp4
Information Unclear	Q: I want to analyze the performance of Manchester United in the Premier League  Correct searching name : Man United
Information Incorrect	Q: "Retrieve the app data of an app with the ID 'com.example.app'.  Result: No such app

Types of Errors (Q: User Query)

# ➤ Problem Analysis of Existing Frameworks: ToolLLaMA

- LLM recognizes when the user query is incomplete and needs clarification.

Query 1: Find a dog age in human years (information missing)

Query 2: Help me find the weather report of the destinations (information missing)

Thought of ToolLLaMA

**Thought:** "To evaluate the expression 'dog age in human years', I need more information. Can you please provide me with the age of dog"

**Thought:** "Ask the user for the names of the destinations they are considering using the function "ask user for destinations"

We can include proactive prompting techniques to handle this situation.



# ➤ Dataset: Itool

New human-annotated benchmark named 'Itool'

- Collects incomplete user queries from four types of errors.
- Existing tool-learning framework cannot handle incomplete queries from users

Case	Num of query
Unclear instruction & instruction with missing information	70
Instruction with problematic searching items	30
Instruction that cannot support	25
Instruction with different meanings	13
Total	138

Data size of Itool





# ➤ Dataset: Itool

- Instruction with missing information

Missing information refers to the absence or lack of necessary details.

## instruction with missing information

```
"query": "When is his latest movie coming out?",  
"question need to be raise": "Who is his ?",  
"answer": "Clint Eastwood",  
"solution": [  
  "GET /search/person",  
  "GET /person/{person_id}/movie_credits",  
  "GET /movie/{movie_id}/release_dates"  
]
```

- Unclear information

Unclear information refers to instances where the user provides information that is not comprehensive enough to complete their query.

## Unclear instruction

```
"query": "I want to know when the movie about the simulation was  
released",  
"question need to be raise": "What does the movie about the simulation  
refer to?",  
"answer": "The Matrix",  
"solution": [  
  "GET /search/movie",  
  "GET /movie/{movie_id}/release_dates"  
]
```



# ➤ # Method: Query when Need (QwN)

## Key ideas of QwN:

- Leverage the ability of LLMs to follow instructions and prompt them to seek the user's clarification whenever necessary  
→ Query when Need (QwN)
- Append user's answer to the tool-augmented models' memory to guide their next step

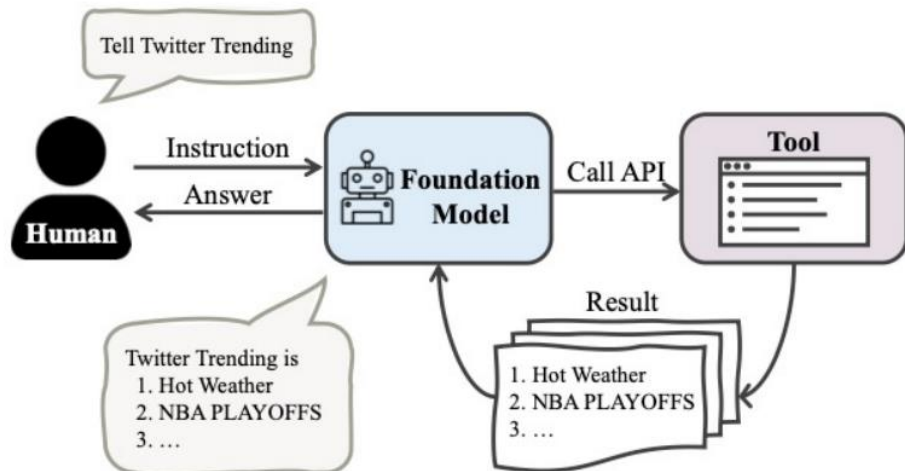
### Proactive Prompting

*Always think comprehensively and seek clarification from users whenever the user's query is not clear enough for you to formulate API calling. If you require additional information or face uncertainty, start the conversation with "I need user's clarification" and then express the challenges explicitly.*

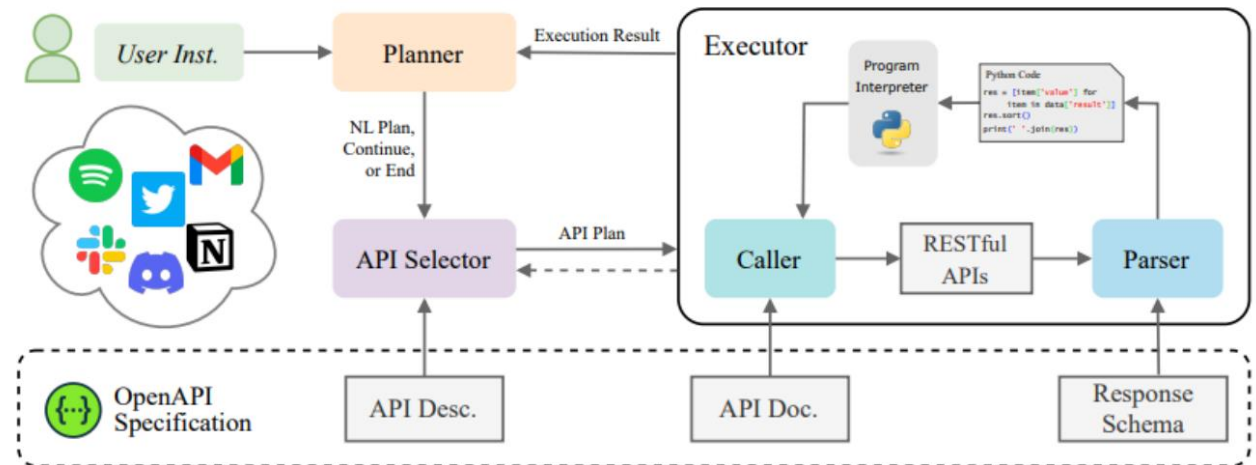
# ➤ # Method: Query when Need (QwN)

## Our research plan:

- Incorporate QwN with the existing tool-augmented frameworks to enhance their ability
- ToolLLM and RestGPT are two representative frameworks



ToolLLaMa (single-agent framework)

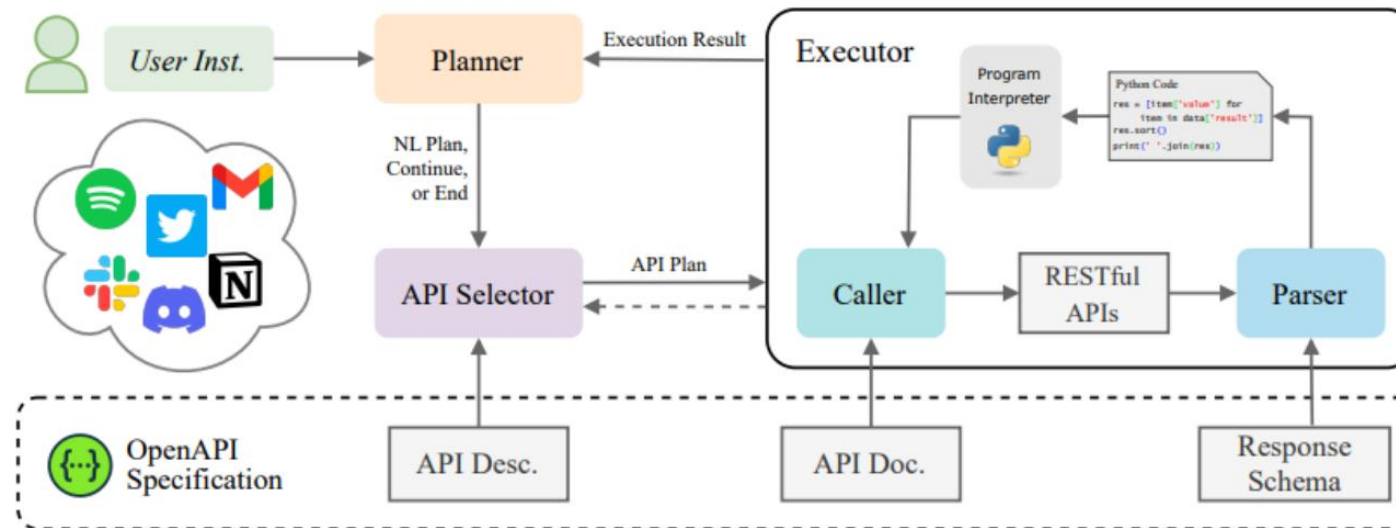


RestGPT (multi-agent framework)

# # RestGPT

## RestGPT:

- Multi-agent framework
- Different agents may encounter different difficulties
- Communicate with single-agent system is easy, how about communicating with multi-agent system?



# ➤ # Method

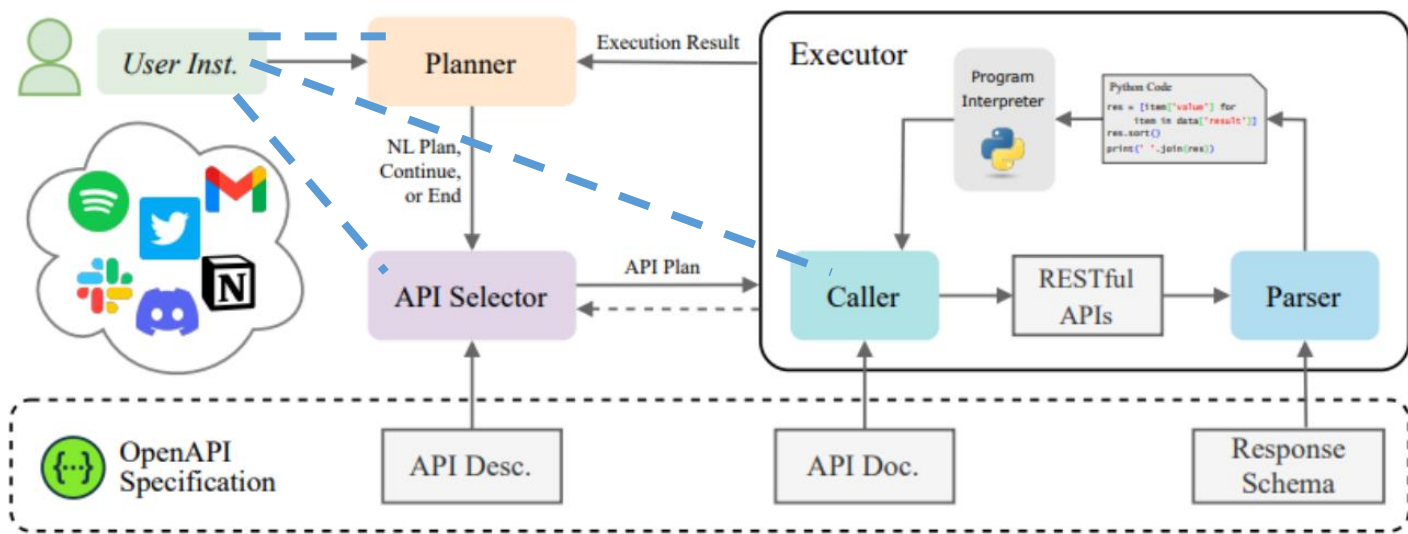


## Most Intuitive Way:

Prompt each agent to communicate with users whenever necessary

## Challenges:

The additional information provided by users cannot be effectively delivered across the framework







## ➤ # Method

Example:

Initial user instruction (Goal of the framework): Search for the father of the director of movie A

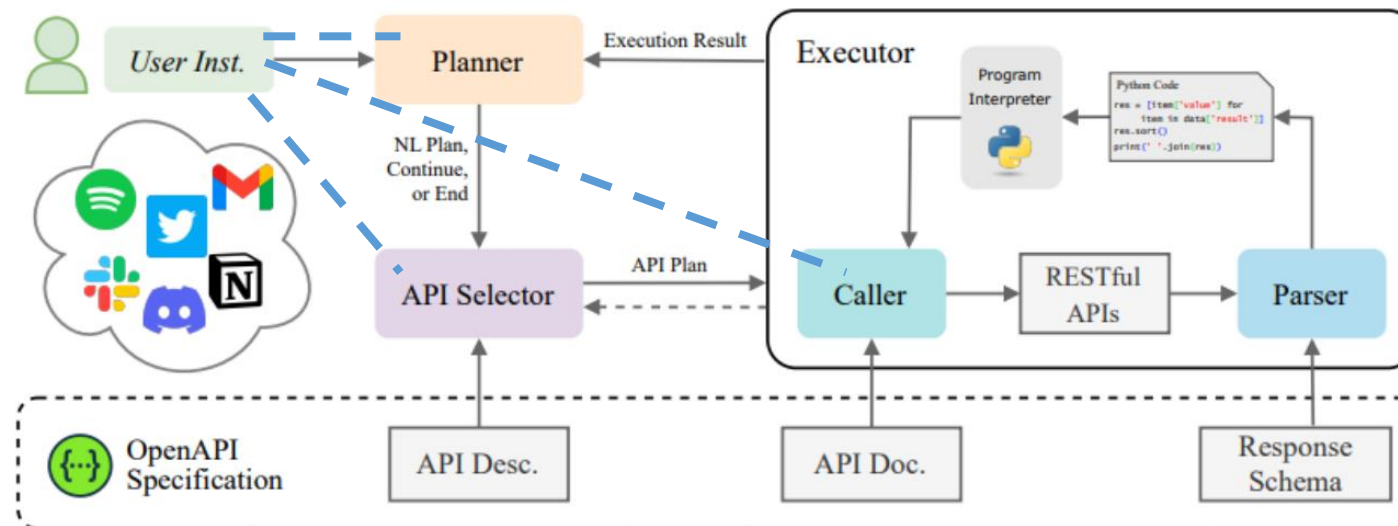
....

Caller may fail to find movie A (if the user did not provide the movie name correctly)

Caller seek help from users and understand that user indeed refer to the movie A' (not A)

Caller can then successfully find movie A'

However, Planner doesn't know the objective of the user has been changed during communication. Therefore, it still responsible for the user's original query, which is to search the father of the director of movie A



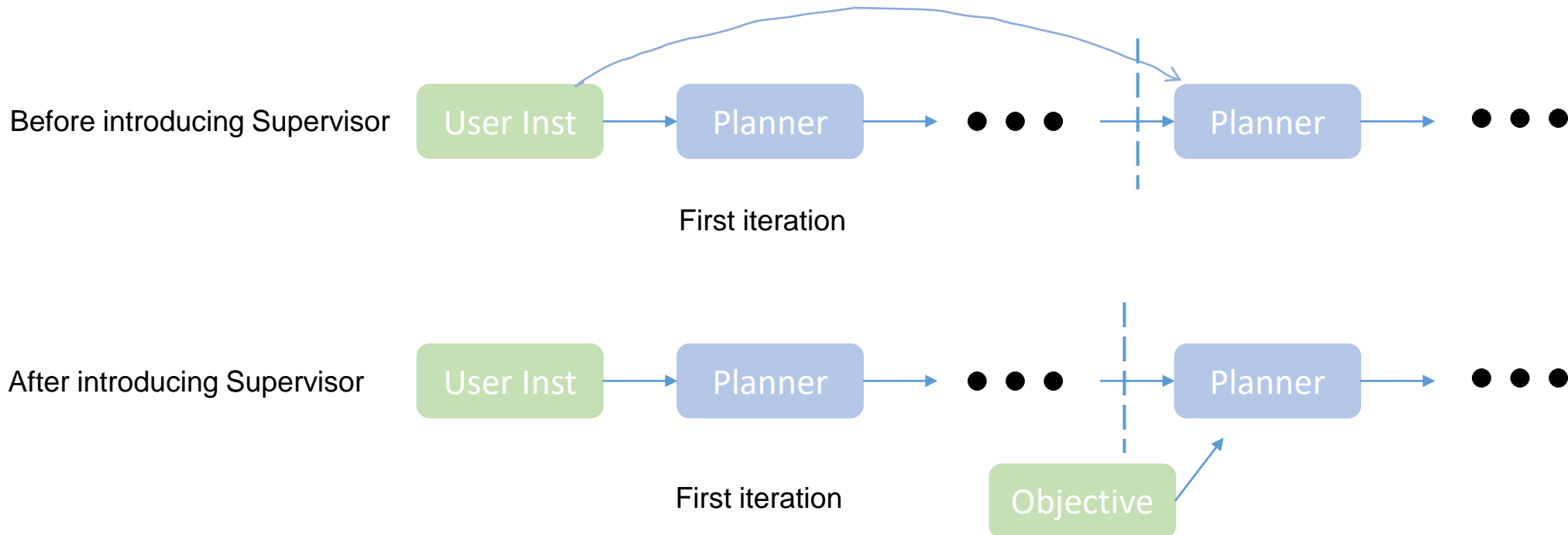
## ➤ # Method



Except for prompting agents to seek help during tool utilization, we introduce a new role Supervisor for the implementation of QwN in multi-agent system

Role of Supervisor:

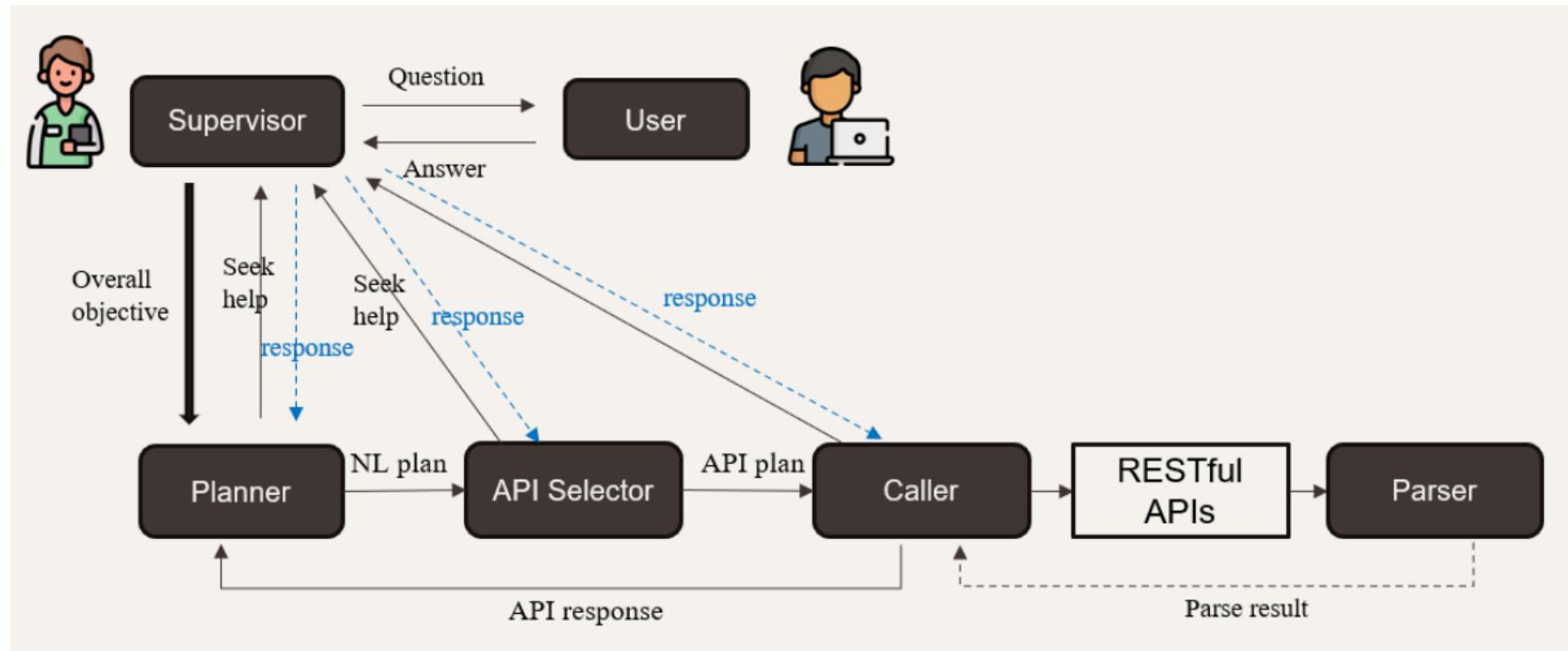
- Hold the global memory of the entire system
- Adjust the overall objective of the Planner after each iteration
- Work as coordinator between the user and the separate agent



# ➤ # Method: QwN-RestGPT

Utilize the global view of Supervisor to do more jobs

- We do not directly allow the agents to communicate with the user. Instead, we use Supervisor as the bridge between users and the separate agents



Why?  
➔

- Avoid raising unnecessary questions
- Analyzing the difficulties can help to raise the desired questions for querying additional information



# # Demo



```
carlos@carlos-virtual-machine: ~/Documents/RestGPT(testi...
carlos@carlos-virtual-machine:~/Documents/RestGPT(testing performance on origina
l dataset)$
```

## # Experiment result



We test Itool on QwN-augmented RestGPT and RestGPT

QwN-augmented RestGPT outperforms RestGPT in:

- Complete instructions
- Unclear or missing information
- Problematic searching item instructions
- Instructions with different meanings

Case	RestGPT	QwN-augmented RestGPT
Complete instruction	0.66 (50 cases)	0.76 (50 cases)
Unclear or missing information	0 (70 cases)	0.61 (70 cases)
Problematic searching items	0.5 (30 cases)	0.93 (30 cases)
Instruction that cannot support	0 (13 cases)	0.31 (13 cases)
Instruction with different meanings	0	-

Experiment result of RestGPT and QwN-augmented RestGPT





## Research Questions

To evaluate QwN-augmented RestGPT, we study three research questions:

- RQ1: Explain that QwN-augmented RestGPT can have a better performance compared to the original RestGPT framework.
- RQ2: Examination of whether QwN-augmented RestGPT can accurately identify the incorrect part in the user query.
- RQ3: Examination of whether QwN-augmented RestGPT can improve the robustness of incomplete queries.



## ➡ Result and Analysis

RQ1: Explain that QwN-augmented RestGPT can have a better performance compared to the original RestGPT framework

- Experimental results of the complete dataset using both QwN-augmented RestGPT and RestGPT
- Supervisor: Adjust the overall objective of the Planner after each iteration

### RestGPT

**Query:** I just finished watching Titanic and I want some other movie recommendations

**Planner:** Continue. | Get the list of **similar** movies to Titanic (597)

**API Selector:** GET /movie/597/**similar** to get the list of similar movies to Titanic (597)

### QwN-augmented RestGPT

**Query:** I just finished watching Titanic and I want some other movie recommendations

**Objective:** Please use the id of the movie "Titanic" (597) to get movie recommendations.

**Planner:** Get the movie recommendations of the movie "Titanic" (597).

**API Selector:** GET /movie/597/recommendations to get the movie recommendations of the movie Titanic (id 597).

# ➤ Result and Analysis



RQ2: Examination of whether QwN-augmented RestGPT can accurately identify the incorrect part in the user query

We label the questions as either ‘relevant’ or ‘irrelevant’

Case	Planner	Caller	Total
Unclear instruction & instruction with missing information			
Relevant	39	16	55
Irrelevant	6	9	15
Instruction with problematic searching items			
Relevant	11	6	17
Irrelevant	2	1	3
Instruction with different meanings			
Relevant	1	2	3
Irrelevant	0	2	2
Total	59	35	95
Relevant: 75 (0.789), Irrelevant: 20 (0.211)			

Irrelevant follow-up question

**Planner:** Search for Breaking Bad (id1) using GET /search/tv.

**API Selector:** Thought: I have the id of Breaking Bad, I can use it to get the details of the show.

**Plan:** GET /tv/id1 to get the details of Breaking Bad.

**Caller:** I need user's clarification. The API response indicates an error: "The resource you requested could not be found." It seems that the id of Breaking Bad is not valid. Please provide the valid id of Breaking Bad.

**Question:** Could you provide the valid id of Breaking Bad?

Number of relevant and irrelevant questions raised up by our model

# ➤ Result and Analysis



RQ3: Examination of whether QwN-augmented RestGPT can improve the robustness of incomplete query

- QwN-augmented RestGPT

## Instructions with missing information

### Instruction of TMDB

**Query:** I need something like that movie.

### QwN-augmented RestGPT

**Planner:** I need user's clarification! The user did not provide enough information to proceed. Without a specific movie title, I cannot search for similar movies.

**Question:** Could you provide the title of the movie you are looking for?

**User:** Titanic

**Additional information:** The movie the user is looking for is Titanic.

**Planner:** Search for a movie with the title "Titanic".

## Problematic Searching Item Instructions

### Instruction of TMDB

**Query:** What is the logo of the Watt Daisney?

### QwN-augmented RestGPT

**Planner:** Search for the company "Watt Daisney".

**Parser:** No results found for the query 'Watt Daisney'.

**Caller:** Execution Result: I need user's clarification. The API response did not return any relevant information when I call GET /search/company to search for the company "Watt Daisney".

**Question:** Could you please double-check if the company name is correct?

**User:** It should be Walt Disney

# ➤ Result and Analysis



## RQ3: Examination of whether QwN-augmented RestGPT can improve the robustness of incomplete query

We also tested our data set on the original RestGPT

```
Instructions with Missing Information
Instruction of TMDB

Query: give me the number of movies directed by my father

RestGPT

Planner: search for the person with name "my father"
API Selector: GET /search/person to search for the person with name
"my father"
Caller: I need to make a GET request to the /search/person endpoint
with the query parameter set to "my father".
Operation: GET Input: {
"url": "https://api.themoviedb.org/3/search/person",
"query": "my father"
}
Parser: No results found
Caller: Thought: I cannot finish executing the plan without knowing some
other information.
Execution Result: Failed to call GET /search/person to search for the
person with name "my father". No results found.
Planner: Continue. search for the person with name "father"
Caller: I should try to search for the person with name "father" instead
of "my father".
Operation: GET Input: {
"url": "https://api.themoviedb.org/3/search/person",
"params": {
"query": "father"
}
}
(Finally fail)
```

Query: give me the number of movies directed by my father

1<sup>st</sup> trial: searching "my father"  
No result found

2<sup>nd</sup> trial: searching "father"  
No result found

Wrong spelling of Catherine Hardwicke

Hallucination of searching item

```
Problematic Searching Item Instructions
Instruction of TMDB

Query: tell me a TV show recently directed by Katherine Hardwood

RestGPT

Planner: search for the director "Katherine Hardwood"
API Selector: GET /search/person to search for Katherine Hardwood
Caller: I need to make a GET request to the /search/person endpoint
with the query parameter set to Katherine Hardwood.
Operation: GET Input: {
"url": "https://api.themoviedb.org/3/search/person",
"params": {
"query": "Katherine Hardwood"
}
}
Parser: No results found for Katherine Hardwood.
Caller: Thought: I cannot finish executing the plan without knowing some
other information.
Execution Result: No results found for Katherine Hardwood.
Planner: search for the director "Kathryn Bigelow"
...
(Finally fail)
```





# ➤ # Limitation & Future work:

## Limitation:

- Despite providing agents with the capability to seek assistance when faced with challenges, determining when and what to ask remains a difficult task for current LLM.
- Hallucination still persists.

## Future work:

- Curate additional diverse datasets that involve various tools
- Incorporate QwN with single-agent tool-augmented framework ToolLLM



**Thanks for listening!**



香港中文大學  
The Chinese University of Hong Kong