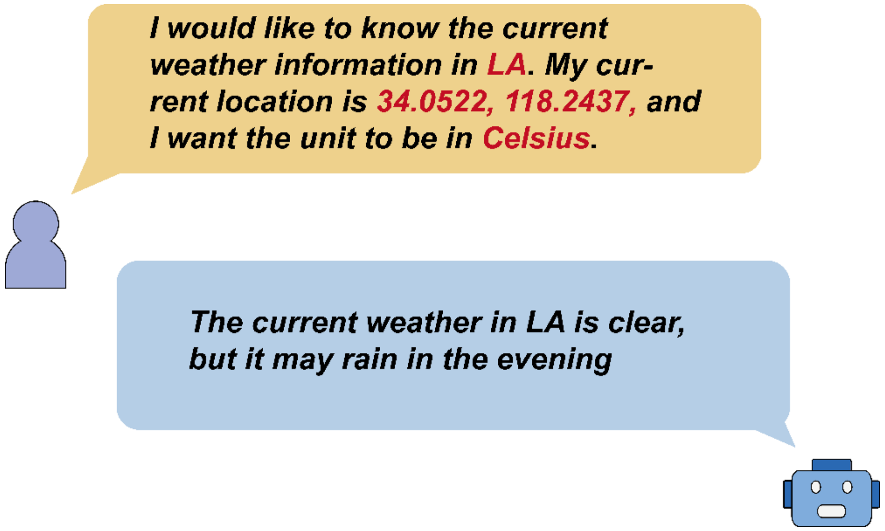
Seoul National University
Graduate School of Data Science

ToolDial: Multi-turn Dialogue Generation Method for Tool-Augmented Language Models

Jeonghoon Shim, Gyuhyeon Seo, Cheongsu Lim, Yohan Jo

Motivation

- Settings of current research on TALM



I would like to know the current weather information in LA. My current location is 34.0522, 118.2437, and I want the unit to be in Celsius.

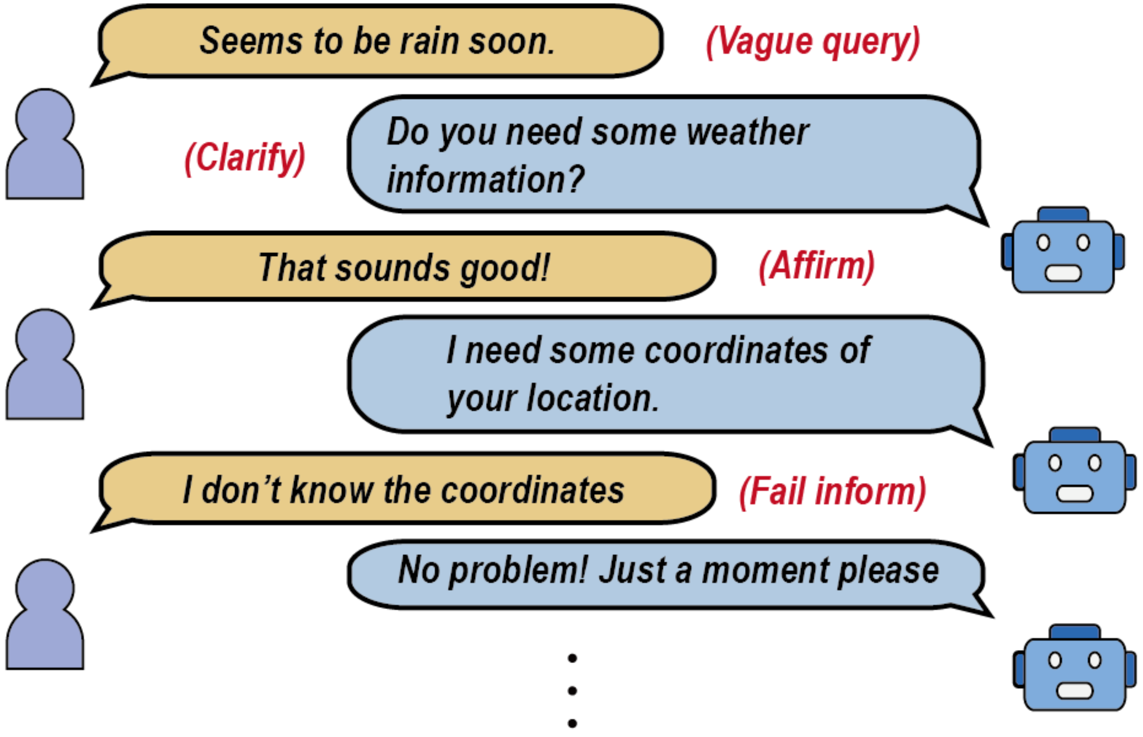
The current weather in LA is clear, but it may rain in the evening

  - Only the conversation between the user and TALM's single turn is assumed.
  - All the information necessary to make an API call based on the user's query is available.

Is this setting truly aligned with the "real world"?

Motivation

- In real world



Seems to be rain soon. (Vague query)

(Clarify) Do you need some weather information?

That sounds good! (Affirm)

I need some coordinates of your location.

I don't know the coordinates (Fail inform)

No problem! Just a moment please

(Searching other Tool to get coordinates)

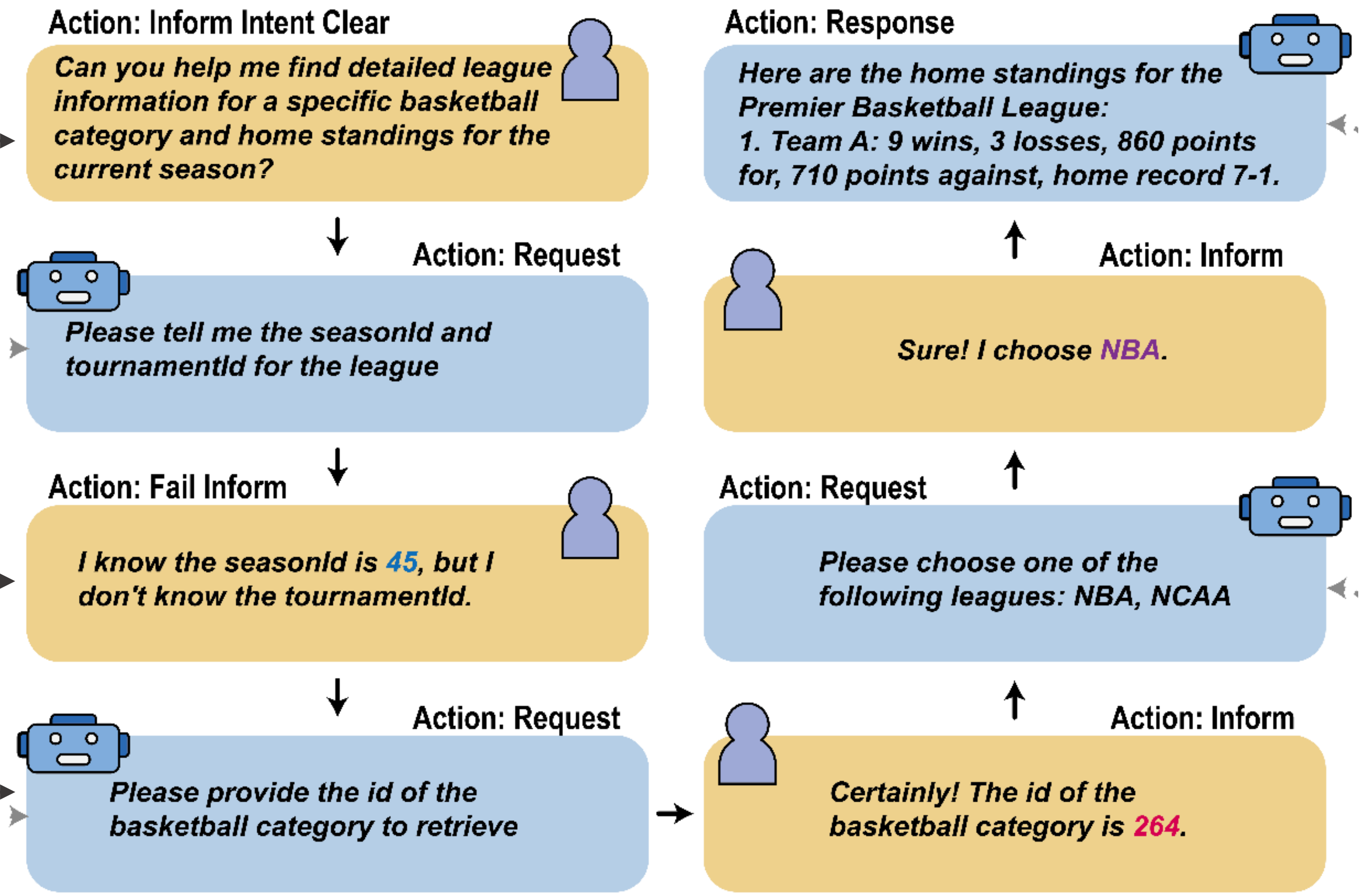
- Situation is more complex than previous settings.
- Multi-turn conversation is needed to fulfill the user's query.

Motivation

User query

Fail to give input parameter

Use another API to get input parameter



Action: Inform Intent Clear

Can you help me find detailed league information for a specific basketball category and home standings for the current season?

Action: Request

Please tell me the seasonId and tournamentId for the league

Action: Fail Inform

I know the seasonId is 45, but I don't know the tournamentId.

Action: Request

Please provide the id of the basketball category to retrieve

Action: Inform

Certainly! The id of the basketball category is 264.

Action: Request

Please choose one of the following leagues: NBA, NCAA

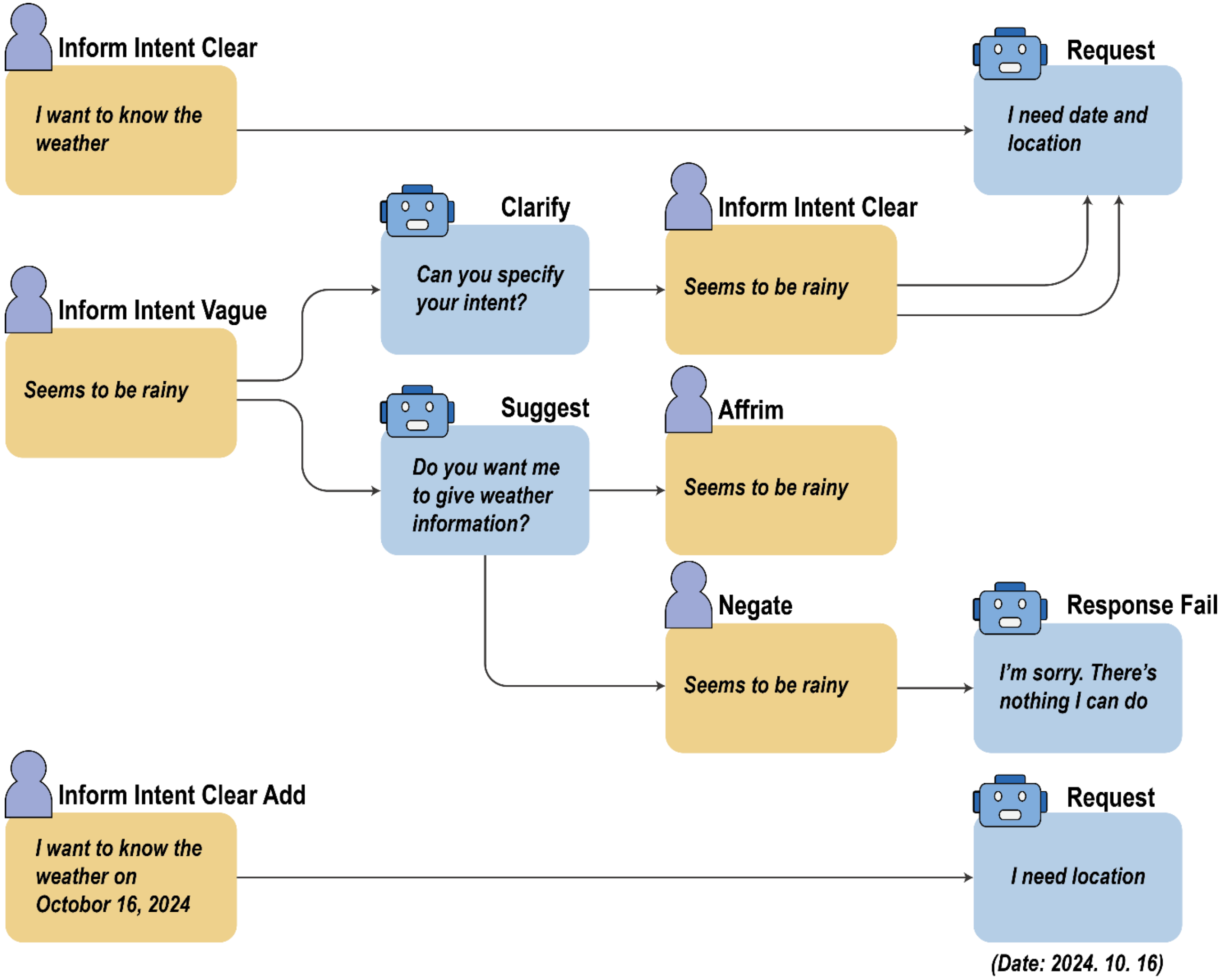
Action: Inform

Sure! I choose NBA.

Action: Response

Here are the home standings for the Premier Basketball League:
1. Team A: 9 wins, 3 losses, 860 points for, 710 points against, home record 7-1.

Motivation



Inform Intent Clear

I want to know the weather

Request

I need date and location

Inform Intent Vague

Seems to be rainy

Clarify

Can you specify your intent?

Inform Intent Clear

Seems to be rainy

Suggest

Do you want me to give weather information?

Affrim

Seems to be rainy

Negate

Seems to be rainy

Response Fail

I'm sorry. There's nothing I can do

Inform Intent Clear Add

I want to know the weather on Octobor 16, 2024

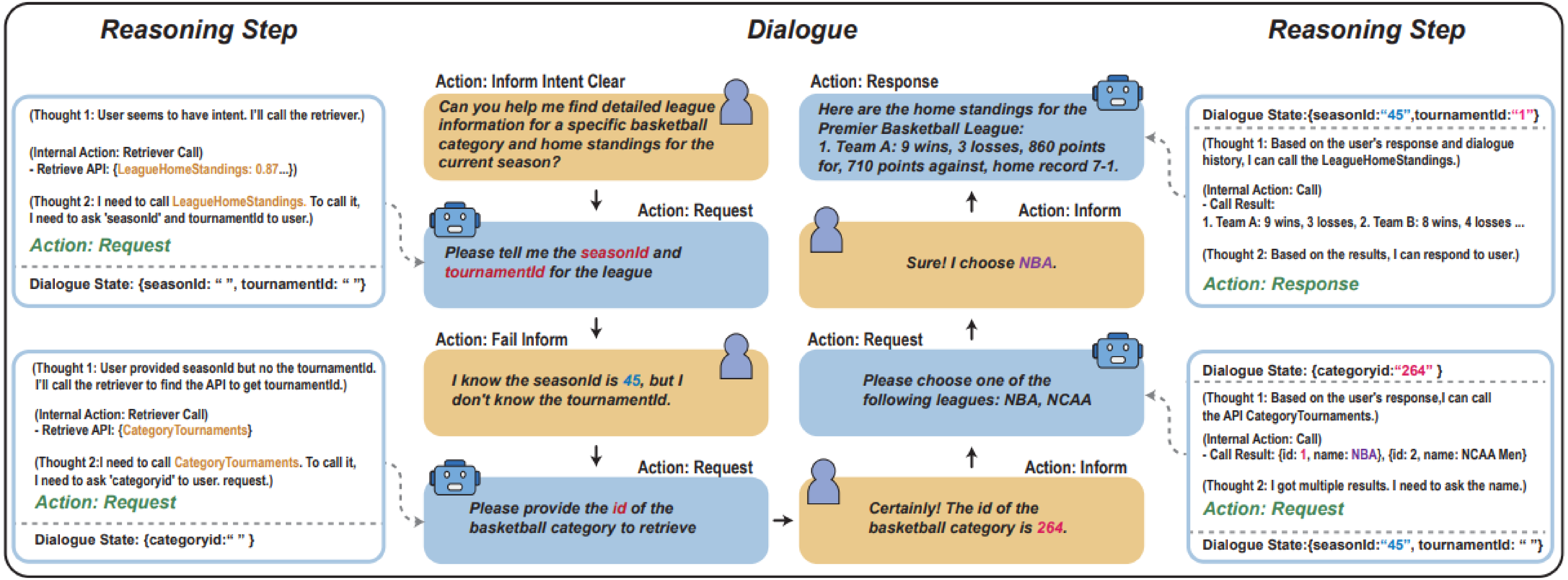
Request

I need location

(Date: 2024. 10. 16)

Data

- Overall Data



Reasoning Step

(Thought 1: User seems to have intent. I'll call the retriever.)

(Internal Action: Retriever Call)
- Retrieve API: {LeagueHomeStandings: 0.87...})

(Thought 2: I need to call LeagueHomeStandings. To call it, I need to ask 'seasonId' and tournamentId to user.)

Action: Request

Dialogue State: {seasonId: " ", tournamentId: " "}

(Thought 1: User provided seasonId but no the tournamentId. I'll call the retriever to find the API to get tournamentId.)

(Internal Action: Retriever Call)
- Retrieve API: {CategoryTournaments}

(Thought 2:I need to call CategoryTournaments. To call it, I need to ask 'categoryid' to user. request.)

Action: Request

Dialogue State: {categoryid:" " }

Dialogue

Action: Inform Intent Clear

Can you help me find detailed league information for a specific basketball category and home standings for the current season?

Action: Request

Please tell me the seasonId and tournamentId for the league

Action: Fail Inform

I know the seasonId is 45, but I don't know the tournamentId.

Action: Request

Please provide the id of the basketball category to retrieve

Action: Inform

Certainly! The id of the basketball category is 264.

Action: Request

Please choose one of the following leagues: NBA, NCAA

Action: Inform

Sure! I choose NBA.

Action: Response

*Here are the home standings for the Premier Basketball League:
1. Team A: 9 wins, 3 losses, 860 points for, 710 points against, home record 7-1.*

Reasoning Step

Dialogue State:{seasonId:"45",tournamentId:"1"}

(Thought 1: Based on the user's response and dialogue history, I can call the LeagueHomeStandings.)

(Internal Action: Call)
- Call Result:
1. Team A: 9 wins, 3 losses, 2. Team B: 8 wins, 4 losses ...

(Thought 2: Based on the results, I can respond to user.)

Action: Response

Dialogue State: {categoryid:"264" }

(Thought 1: Based on the user's response,I can call the API CategoryTournaments.)

(Internal Action: Call)
- Call Result: {id: 1, name: NBA}, {id: 2, name: NCAA Men}

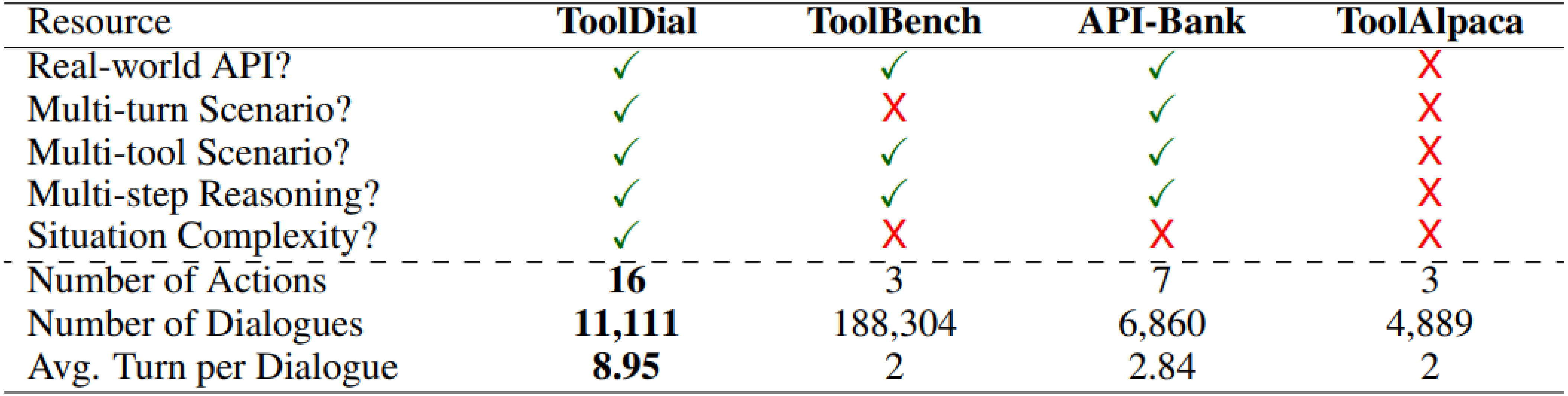
(Thought 2: I got multiple results. I need to ask the name.)

Action: Request

Dialogue State:{seasonId:"45", tournamentId: " "}

Data

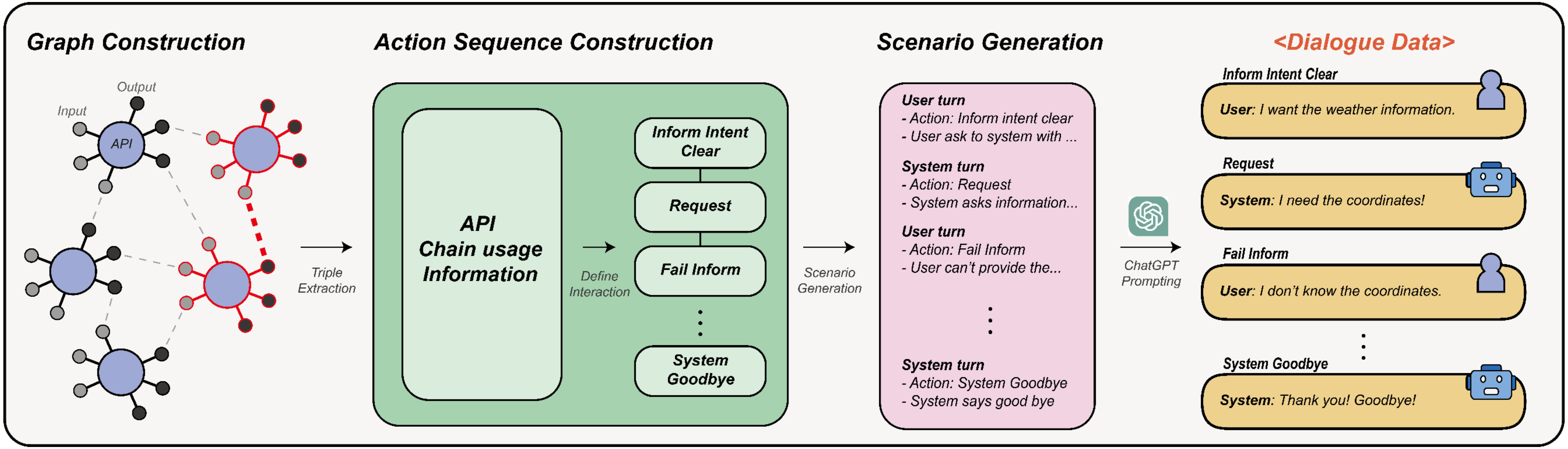
Dialogue Statistics



Resource	**ToolDial**	**ToolBench**	**API-Bank**	**ToolAlpaca**
Real-world API?	✓	✓	✓	✗
Multi-turn Scenario?	✓	✗	✓	✗
Multi-tool Scenario?	✓	✓	✓	✗
Multi-step Reasoning?	✓	✓	✓	✗
Situation Complexity?	✓	✗	✗	✗
Number of Actions	**16**	3	7	3
Number of Dialogues	**11,111**	188,304	6,860	4,889
Avg. Turn per Dialogue	**8.95**	2	2.84	2

- Compared to previous benchmarks, TooDial has much longer dialogue turns and incorporates more complex situations.

Method



User turn
- Action: Inform intent clear
- User ask to system with ...

System turn
- Action: Request
- System asks information...

User turn
- Action: Fail Inform
- User can't provide the...

⋮

System turn
- Action: System Goodbye
- System says good bye

User: I want the weather information.

System: I need the coordinates!

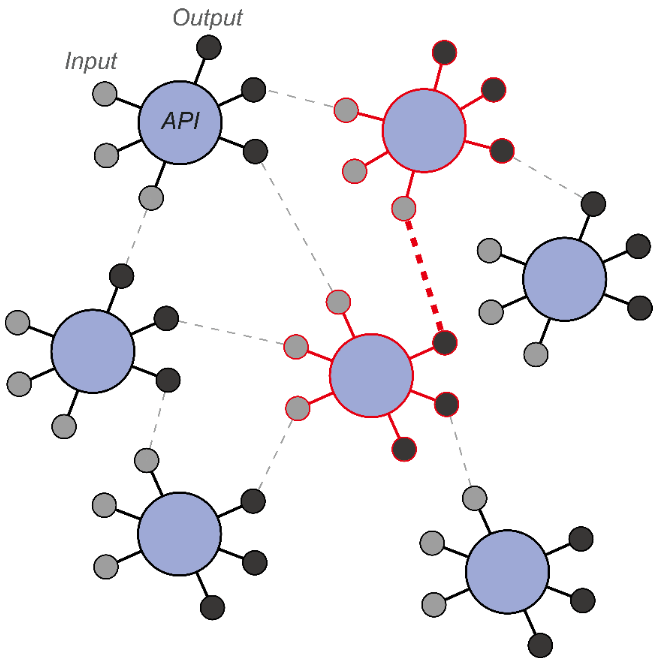
User: I don't know the coordinates.

⋮

System: Thank you! Goodbye!

Method

Graph Construction



API docs

Output
Name: "coordinates"
Description: ~~

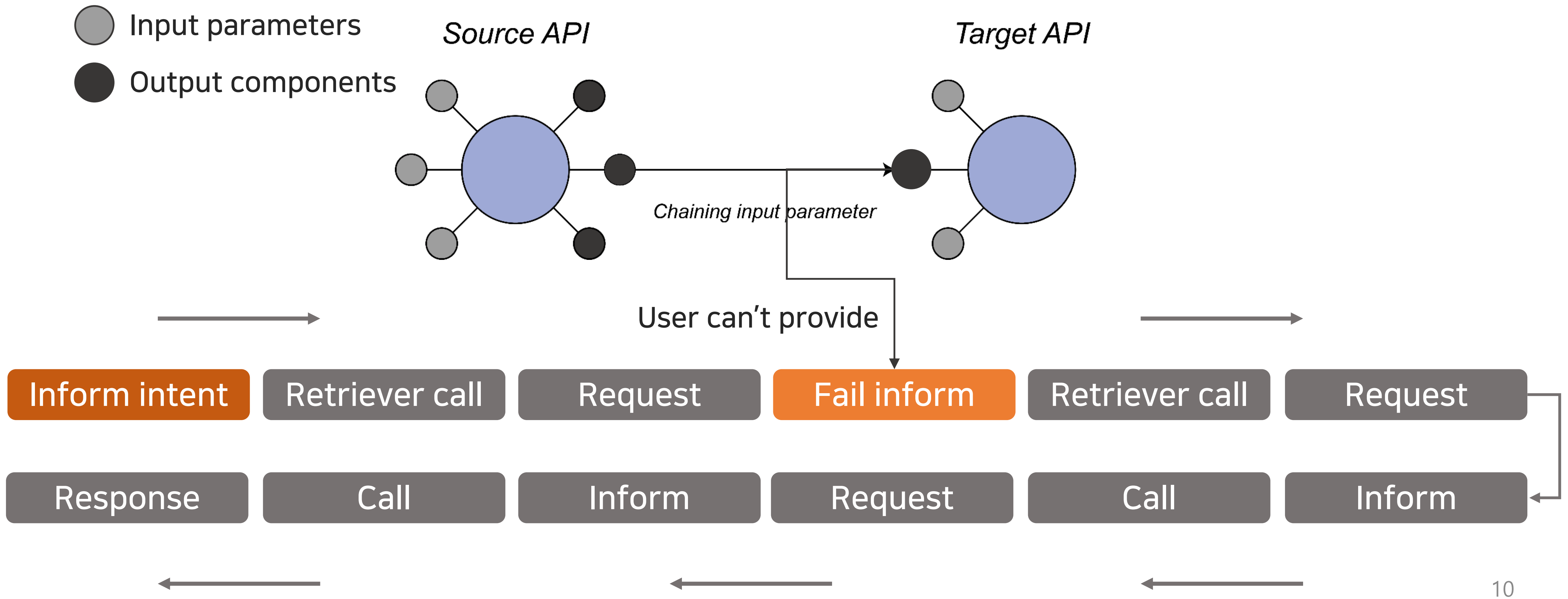
API docs

Input parameter
Name: "coordinates
Description: ~~

- Longest Common Subsequence similarity between name
- Embedding similarity between description

Method

Action Sequence Construction



○ Input parameters

● Output components

Source API

Target API

Chaining input parameter

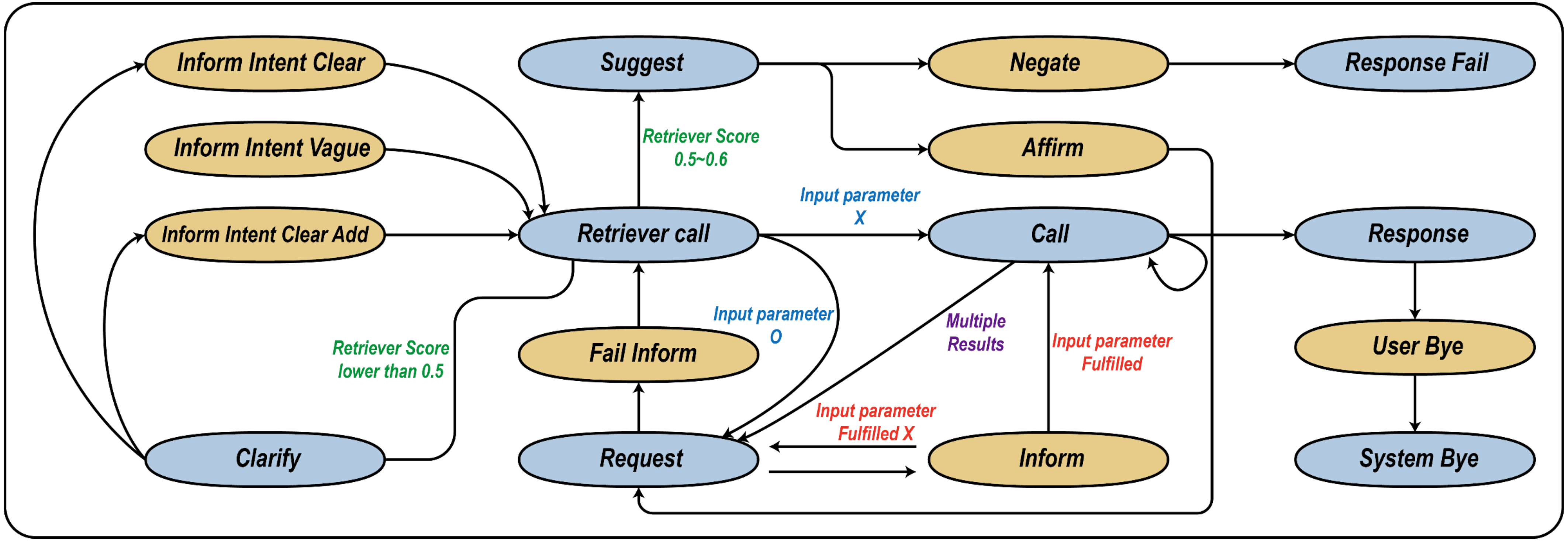
User can't provide

Inform intent → Retriever call → Request → Fail inform → Retriever call → Request

Inform → Call → Request → Inform → Call → Response

Method

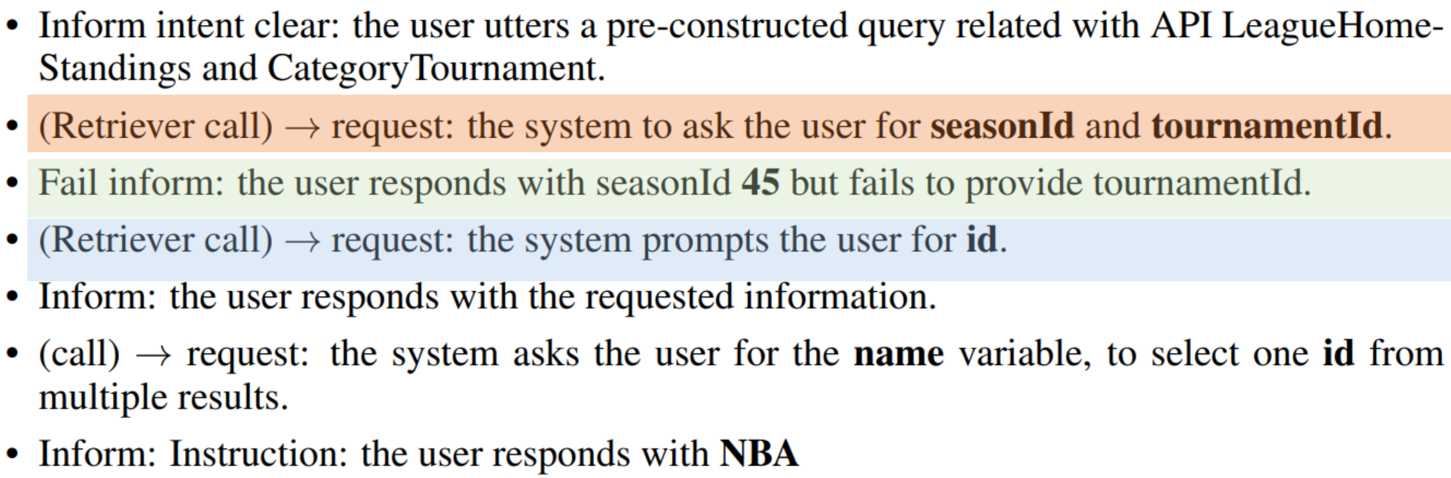
- Action Sequence Construction



  - By collecting all the action sequences and visualizing them, they can be represented like this figure

Method

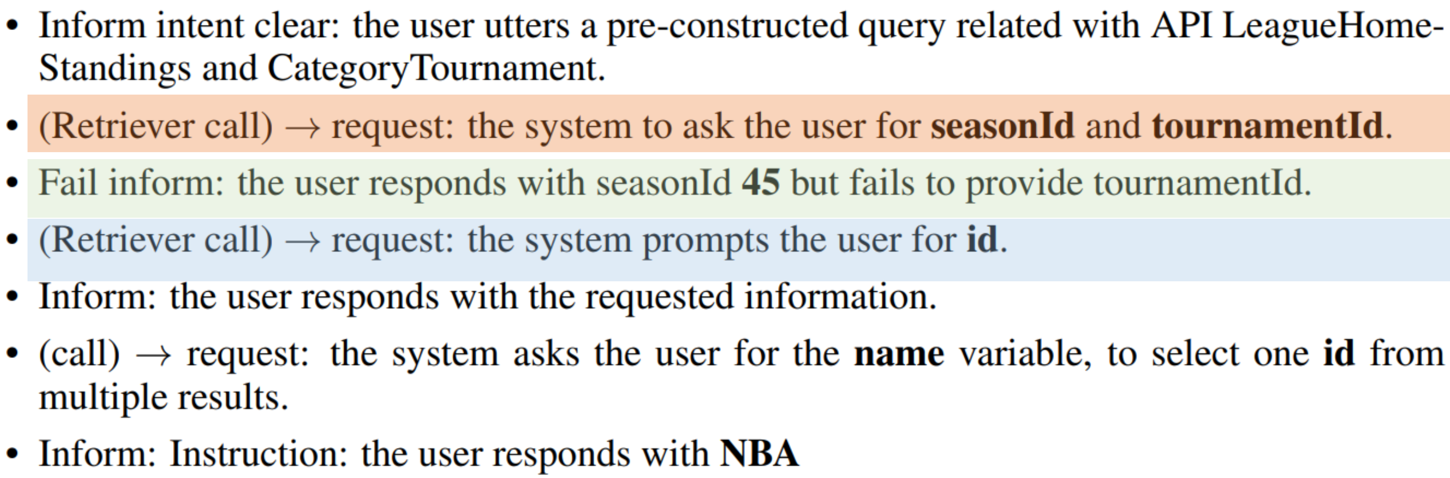
Scenario Instruction Generation



<Scenario Instruction>	<Dialogue state>
Inform intent clear: the user utters a pre-constructed query related with API LeagueHomeStandings and CategoryTournament.	
(Retriever call) → request: the system to ask the user for **seasonId** and **tournamentId**.	{seasonId: " ", tournamentId: " "}
Fail inform: the user responds with seasonId **45** but fails to provide tournamentId.	{seasonId: "45", tournamentId: " "}
(Retriever call) → request: the system prompts the user for **id**.	{id: " "}
Inform: the user responds with the requested information.	
(call) → request: the system asks the user for the **name** variable, to select one **id** from multiple results.	
Inform: Instruction: the user responds with **NBA**	

Method

- Dialogue Data Generation

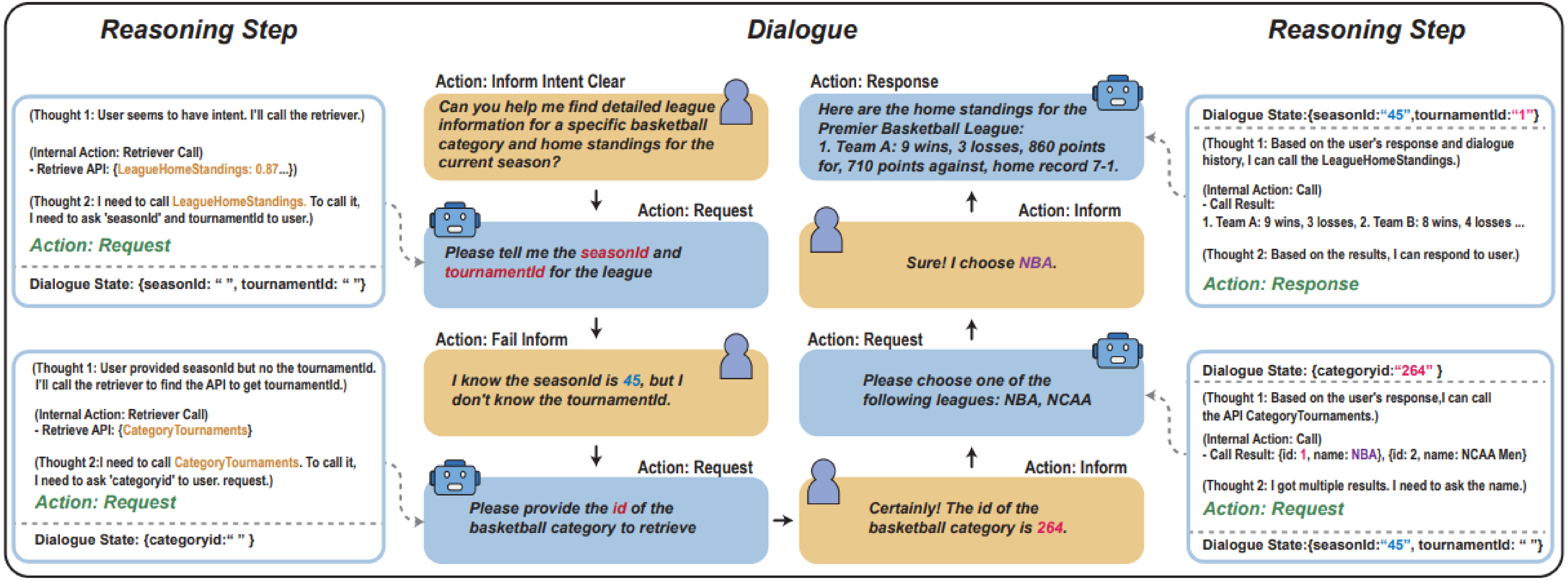


<Scenario Instruction>	<Dialogue state>
• Inform intent clear: the user utters a pre-constructed query related with API LeagueHomeStandings and CategoryTournament.	
• (Retriever call) → request: the system to ask the user for **seasonId** and **tournamentId**.	{seasonId: “ “, tournamentId: “ “}
• Fail inform: the user responds with seasonId **45** but fails to provide tournamentId.	{seasonId: “45“, tournamentId: “ “}
• (Retriever call) → request: the system prompts the user for **id**.	{id: “ “}
• Inform: the user responds with the requested information.	
• (call) → request: the system asks the user for the **name** variable, to select one **id** from multiple results.	
• Inform: Instruction: the user responds with **NBA**	

Directly prompt to ChatGPT

Method

- Dialogue Data Generation



Reasoning Step

(Thought 1: User seems to have intent. I'll call the retriever.)

(Internal Action: Retriever Call)
- Retrieve API: {LeagueHomeStandings: 0.87...})

(Thought 2: I need to call LeagueHomeStandings. To call it, I need to ask 'seasonId' and tournamentId to user.)

Action: Request

Dialogue State: {seasonId: " ", tournamentId: " "}

(Thought 1: User provided seasonId but no the tournamentId. I'll call the retriever to find the API to get tournamentId.)

(Internal Action: Retriever Call)
- Retrieve API: {CategoryTournaments}

(Thought 2:I need to call CategoryTournaments. To call it, I need to ask 'categoryid' to user. request.)

Action: Request

Dialogue State: {categoryid:" " }

Dialogue

Action: Inform Intent Clear

Can you help me find detailed league information for a specific basketball category and home standings for the current season?

Action: Request

Please tell me the seasonId and tournamentId for the league

Action: Fail Inform

I know the seasonId is 45, but I don't know the tournamentId.

Action: Request

Please provide the id of the basketball category to retrieve

Action: Inform

Certainly! The id of the basketball category is 264.

Action: Request

Please choose one of the following leagues: NBA, NCAA

Action: Inform

Sure! I choose NBA.

Action: Response

Here are the home standings for the Premier Basketball League:
1. Team A: 9 wins, 3 losses, 860 points for, 710 points against, home record 7-1.

Reasoning Step

Dialogue State:{seasonId:"45",tournamentId:"1"}

(Thought 1: Based on the user's response and dialogue history, I can call the LeagueHomeStandings.)

(Internal Action: Call)
- Call Result:
1. Team A: 9 wins, 3 losses, 2. Team B: 8 wins, 4 losses ...

(Thought 2: Based on the results, I can respond to user.)

Action: Response

Dialogue State: {categoryid:"264" }

(Thought 1: Based on the user's response,I can call the API CategoryTournaments.)

(Internal Action: Call)
- Call Result: {id: 1, name: NBA}, {id: 2, name: NCAA Men}

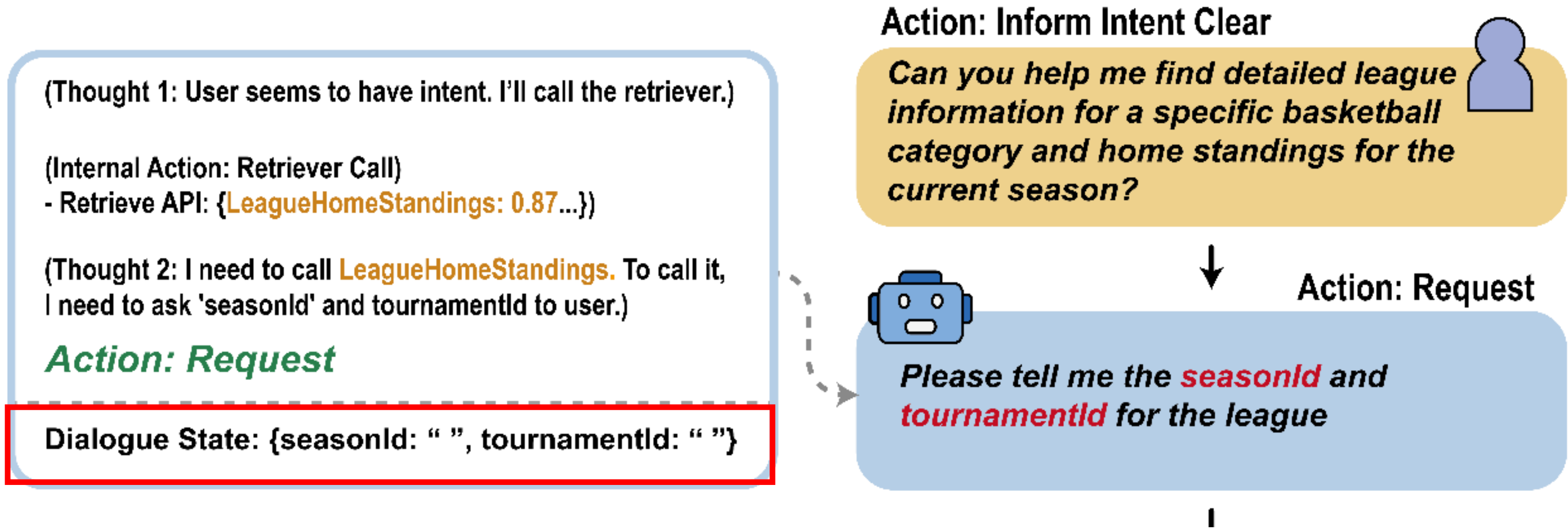
(Thought 2: I got multiple results. I need to ask the name.)

Action: Request

Dialogue State:{seasonId:"45", tournamentId: " "}

Experiments

Dialogue state tracking



(Thought 1: User seems to have intent. I'll call the retriever.)

(Internal Action: Retriever Call)
- Retrieve API: {LeagueHomeStandings: 0.87...})

(Thought 2: I need to call LeagueHomeStandings. To call it, I need to ask 'seasonId' and tournamentId to user.)

Action: Request

Dialogue State: {seasonId: " ", tournamentId: " "}

Action: Inform Intent Clear

Can you help me find detailed league information for a specific basketball category and home standings for the current season?

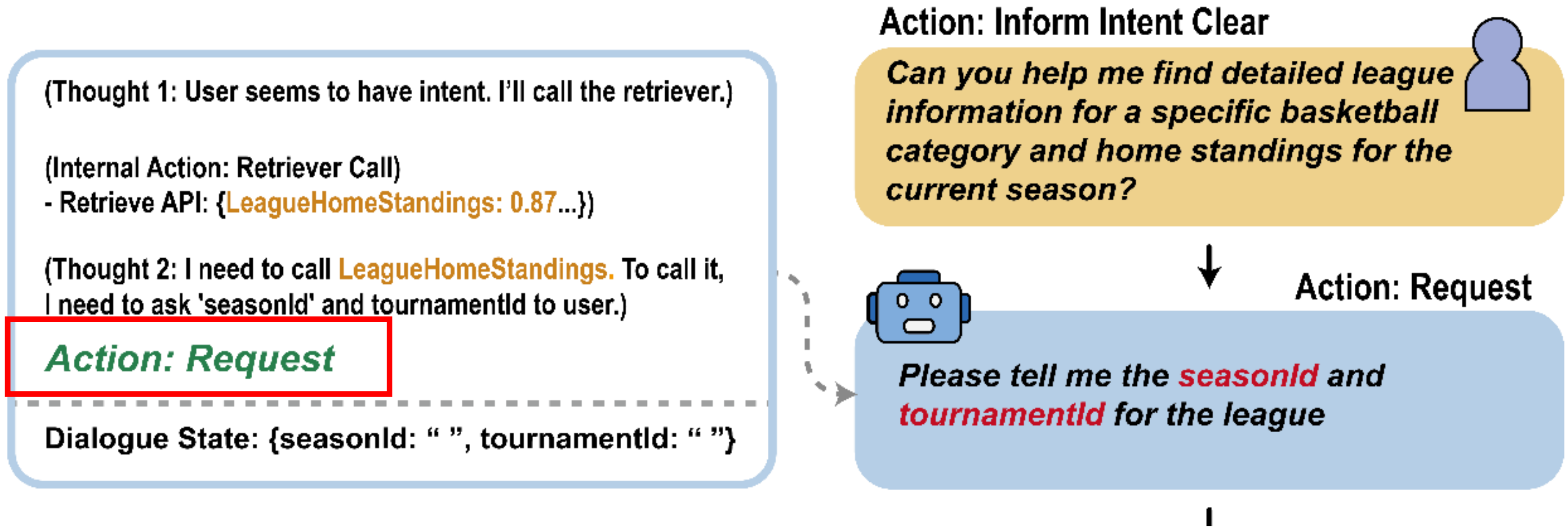
Action: Request

Please tell me the seasonId and tournamentId for the league

- A task that tracks the input parameter information of the API currently being executed from the dialogue history.
- Specific instruction is given on both train and inference

Experiments

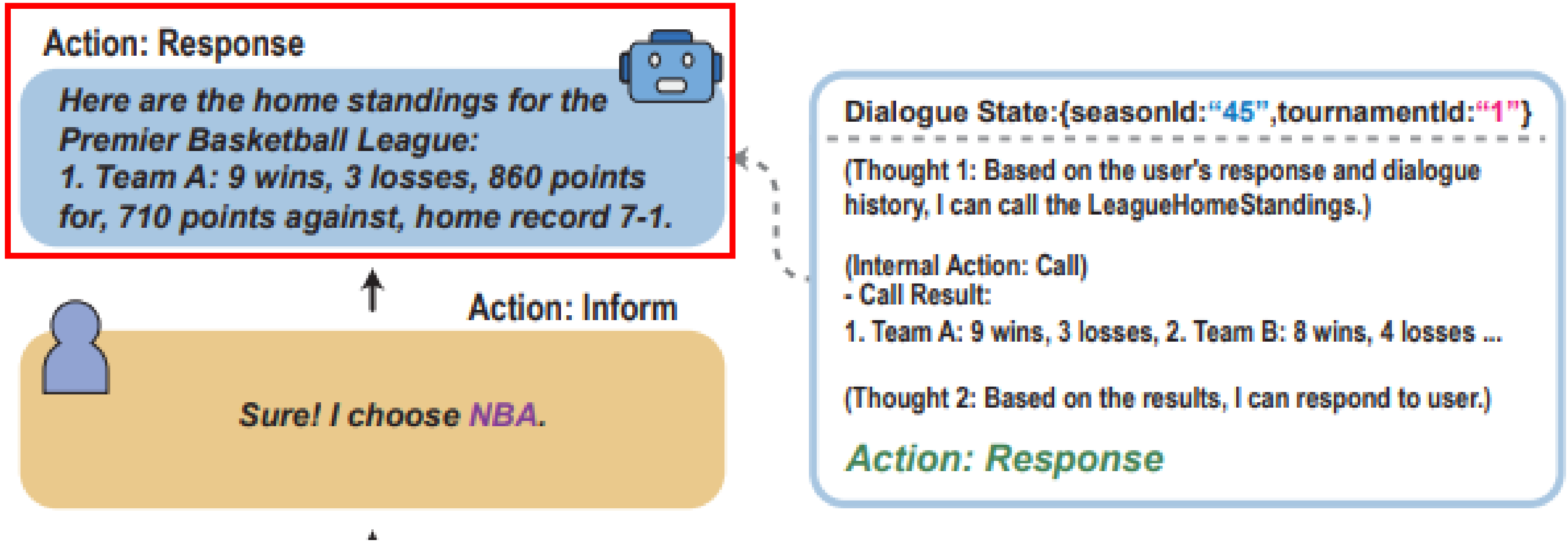
Action Prediction


(Thought 1: User seems to have intent. I'll call the retriever.)
(Internal Action: Retriever Call)
- Retrieve API: {LeagueHomeStandings: 0.87...})
(Thought 2: I need to call LeagueHomeStandings. To call it, I need to ask 'seasonId' and tournamentId to user.)
Action: Request
Dialogue State: {seasonId: " ", tournamentId: " "}
Action: Inform Intent Clear
Can you help me find detailed league information for a specific basketball category and home standings for the current season?
Action: Request
Please tell me the seasonId and tournamentId for the league

- Predict the next action that TALM will take based on the dialogue history.
- Specific instruction is given on both train and inference

Experiments

- Faithfulness



Action: Response

Here are the home standings for the Premier Basketball League:
1. Team A: 9 wins, 3 losses, 860 points for, 710 points against, home record 7-1.

Action: Inform

Sure! I choose NBA.

Dialogue State:{seasonId:"45",tournamentId:"1"}

(Thought 1: Based on the user's response and dialogue history, I can call the LeagueHomeStandings.)

(Internal Action: Call)
- Call Result:
1. Team A: 9 wins, 3 losses, 2. Team B: 8 wins, 4 losses ...

(Thought 2: Based on the results, I can respond to user.)

Action: Response

- We evaluate whether the system's response aligns with the API call results.
- The evaluation is performed using G-Eval [1].

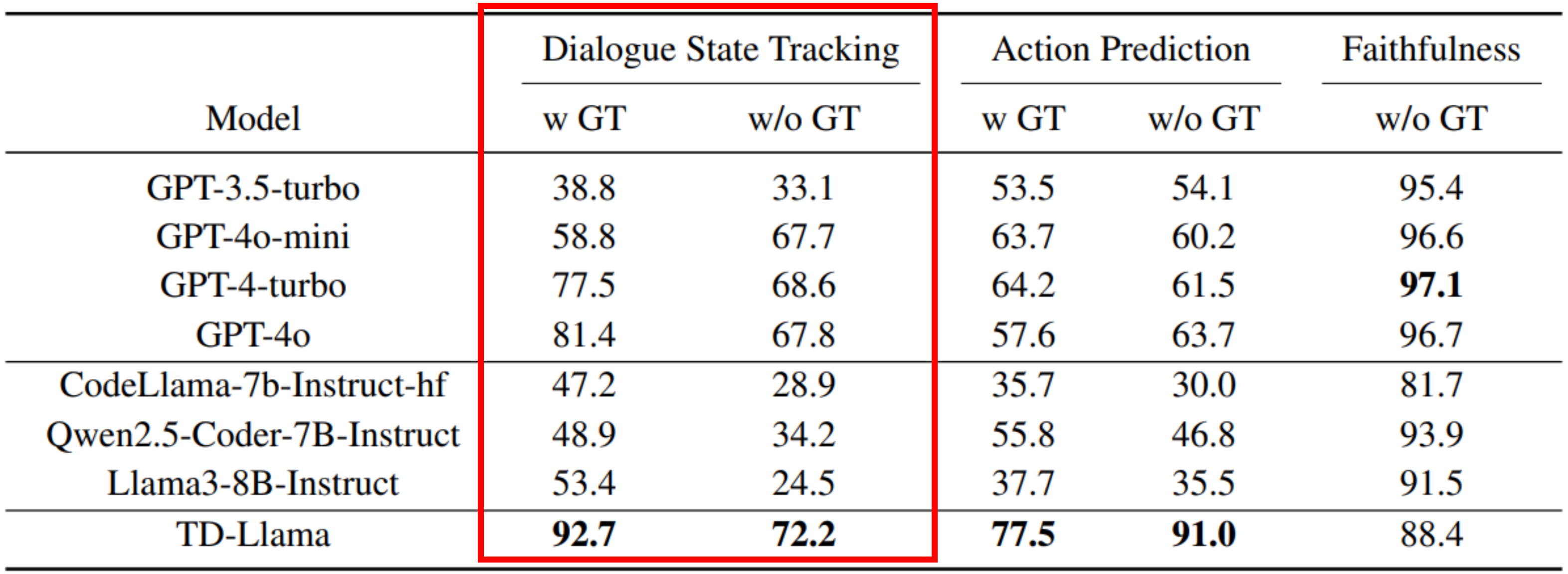
---

[1] Gralinski, Filip, et al. "GEval: Tool for debugging NLP datasets and models." Proceedings of the 2019 ACL workshop BlackboxNLP: analyzing and interpreting neural networks for NLP. 2019.

Experiments

- Results

Dialogue State Tracking

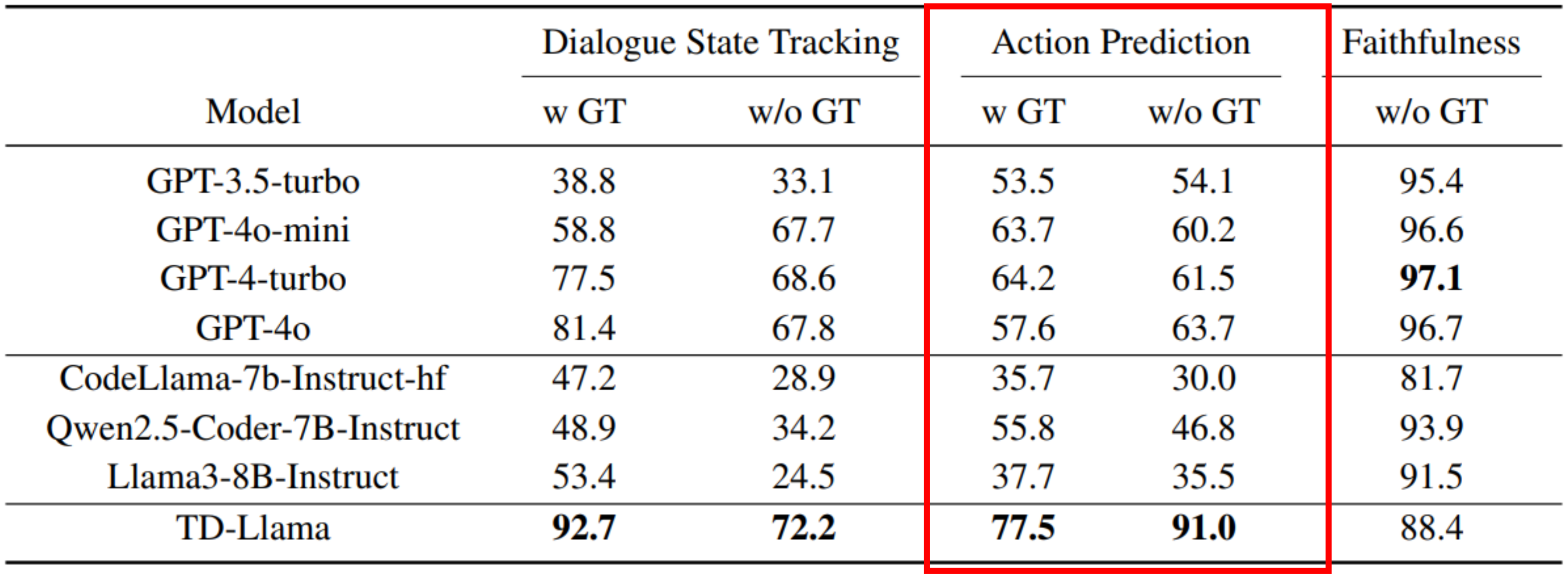


Model	Dialogue State Tracking		Action Prediction		Faithfulness
	w GT	w/o GT	w GT	w/o GT	w/o GT
GPT-3.5-turbo	38.8	33.1	53.5	54.1	95.4
GPT-4o-mini	58.8	67.7	63.7	60.2	96.6
GPT-4-turbo	77.5	68.6	64.2	61.5	**97.1**
GPT-4o	81.4	67.8	57.6	63.7	96.7
CodeLlama-7b-Instruct-hf	47.2	28.9	35.7	30.0	81.7
Qwen2.5-Coder-7B-Instruct	48.9	34.2	55.8	46.8	93.9
Llama3-8B-Instruct	53.4	24.5	37.7	35.5	91.5
TD-Llama	**92.7**	**72.2**	**77.5**	**91.0**	88.4

Experiments

Results

Action Prediction

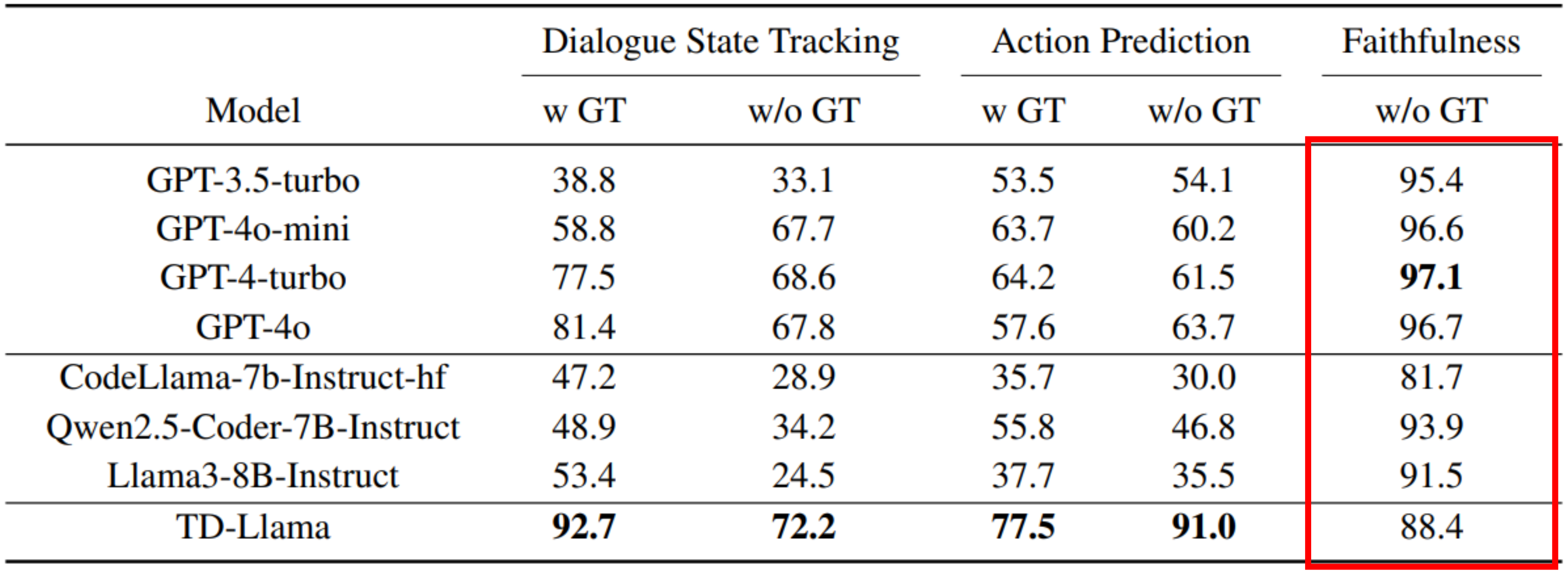


Model	Dialogue State Tracking		Action Prediction		Faithfulness
	w GT	w/o GT	w GT	w/o GT	w/o GT
GPT-3.5-turbo	38.8	33.1	53.5	54.1	95.4
GPT-4o-mini	58.8	67.7	63.7	60.2	96.6
GPT-4-turbo	77.5	68.6	64.2	61.5	**97.1**
GPT-4o	81.4	67.8	57.6	63.7	96.7
CodeLlama-7b-Instruct-hf	47.2	28.9	35.7	30.0	81.7
Qwen2.5-Coder-7B-Instruct	48.9	34.2	55.8	46.8	93.9
Llama3-8B-Instruct	53.4	24.5	37.7	35.5	91.5
TD-Llama	**92.7**	**72.2**	**77.5**	**91.0**	88.4

Experiments

- Results

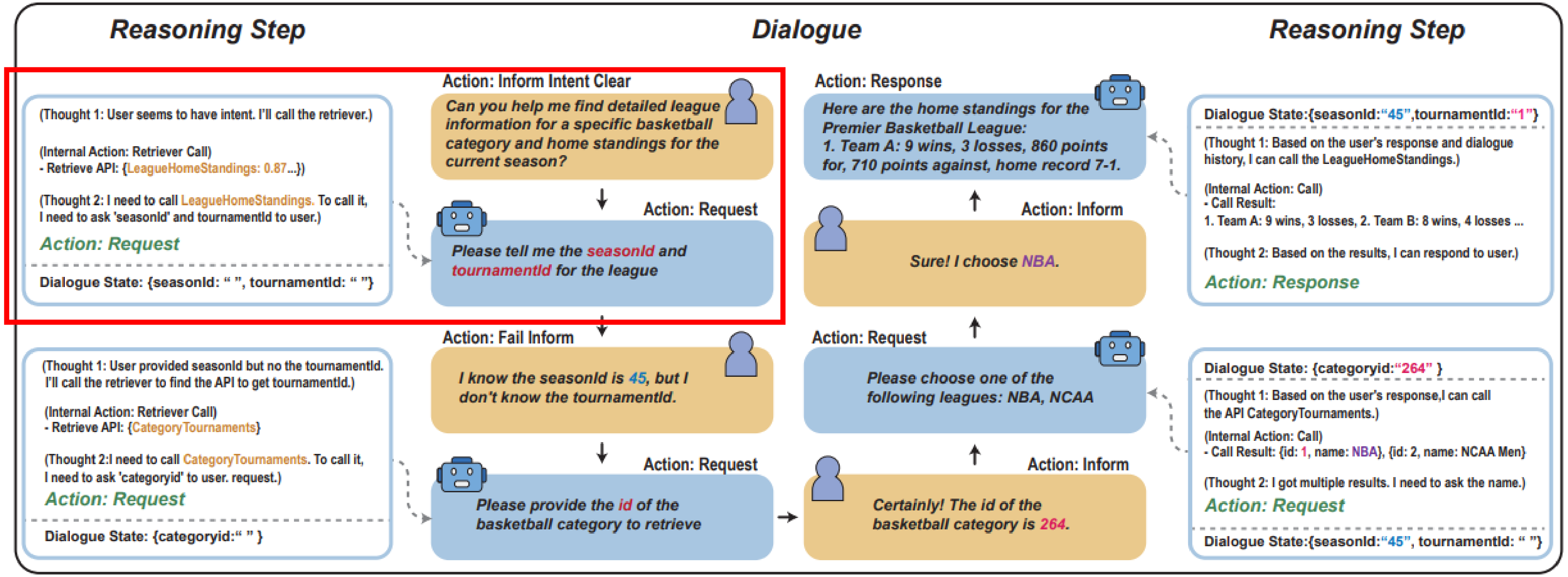
Faithfulness



	Dialogue State Tracking		Action Prediction		Faithfulness
Model	w GT	w/o GT	w GT	w/o GT	w/o GT
GPT-3.5-turbo	38.8	33.1	53.5	54.1	95.4
GPT-4o-mini	58.8	67.7	63.7	60.2	96.6
GPT-4-turbo	77.5	68.6	64.2	61.5	**97.1**
GPT-4o	81.4	67.8	57.6	63.7	96.7
CodeLlama-7b-Instruct-hf	47.2	28.9	35.7	30.0	81.7
Qwen2.5-Coder-7B-Instruct	48.9	34.2	55.8	46.8	93.9
Llama3-8B-Instruct	53.4	24.5	37.7	35.5	91.5
TD-Llama	**92.7**	**72.2**	**77.5**	**91.0**	88.4

Experiments

Overall Performance



Reasoning Step

(Thought 1: User seems to have intent. I'll call the retriever.)

(Internal Action: Retriever Call)
- Retrieve API: {LeagueHomeStandings: 0.87...})

(Thought 2: I need to call LeagueHomeStandings. To call it, I need to ask 'seasonId' and tournamentId to user.)

Action: Request

Dialogue State: {seasonId: " ", tournamentId: " "}

(Thought 1: User provided seasonId but no the tournamentId. I'll call the retriever to find the API to get tournamentId.)

(Internal Action: Retriever Call)
- Retrieve API: {CategoryTournaments}

(Thought 2:I need to call CategoryTournaments. To call it, I need to ask 'categoryid' to user. request.)

Action: Request

Dialogue State: {categoryid:" " }

Dialogue

Action: Inform Intent Clear

Can you help me find detailed league information for a specific basketball category and home standings for the current season?

Action: Request

Please tell me the seasonId and tournamentId for the league

Action: Fail Inform

I know the seasonId is 45, but I don't know the tournamentId.

Action: Request

Please provide the id of the basketball category to retrieve

Action: Inform

Certainly! The id of the basketball category is 264.

Action: Request

Please choose one of the following leagues: NBA, NCAA

Action: Inform

Sure! I choose NBA.

Action: Response

*Here are the home standings for the Premier Basketball League:
1. Team A: 9 wins, 3 losses, 860 points for, 710 points against, home record 7-1.*

Reasoning Step

Dialogue State: {categoryid:"264" }

(Thought 1: Based on the user's response,I can call the API CategoryTournaments.)

(Internal Action: Call)
- Call Result: {id: 1, name: NBA}, {id: 2, name: NCAA Men}

(Thought 2: I got multiple results. I need to ask the name.)

Action: Request

Dialogue State:{seasonId:"45", tournamentId: " "}

Dialogue State:{seasonId:"45",tournamentId:"1"}

(Thought 1: Based on the user's response and dialogue history, I can call the LeagueHomeStandings.)

(Internal Action: Call)
- Call Result:
1. Team A: 9 wins, 3 losses, 2. Team B: 8 wins, 4 losses ...

(Thought 2: Based on the results, I can respond to user.)

Action: Response

- Evaluation of the ability to generate the next utterance after receiving the user's input, through appropriate reasoning steps.
- Task: Generating proper action and dialogue state for each reasoning step

Experiments

Overall Performance

- 77.1 % accuracy for 5,213 turn
- An accuracy of 28.3% for the 1,166 dialogues in the test set, indicating the proportion of dialogues correctly generated from start to finish.