LLMs and Tools
Function Calling

Advanced Large Language Models

ELL8299 · AIL861 · ELL881

Dinesh Raghu
Senior Researcher, IBM Research

LLMs and Tools

Part 1: Incorporating Tools during Fine-tuning (Tool Augmentation)

Part 2: Teaching LLMs to Use External APIs (Function Calling)

Part 3: Automating Complex Tasks (AI Agents)

Recap

1. APIBench
   - Single turn dialogs grounded on single tool

2. ToolAlpaca
   - Multi turn dialogs grounded on single tool

3. **ToolBench**
   - **Multi turn dialogs grounded on multiple tools**

4. xlam-function-calling-60k
   - Single turn dialogs grounded on multiple tool

Research Directions

High Fidelity Data Synthesis

- TOOLFLOW
- MAGNET

Beyond SFT: RL-Enhanced Finetuning

- MAGNET

Towards Realistic Evaluation

- TAU (τ) – Bench
- τ ^{2} - Bench

Research Directions

High Fidelity Data Synthesis

- ToolFlow
- MAGNET

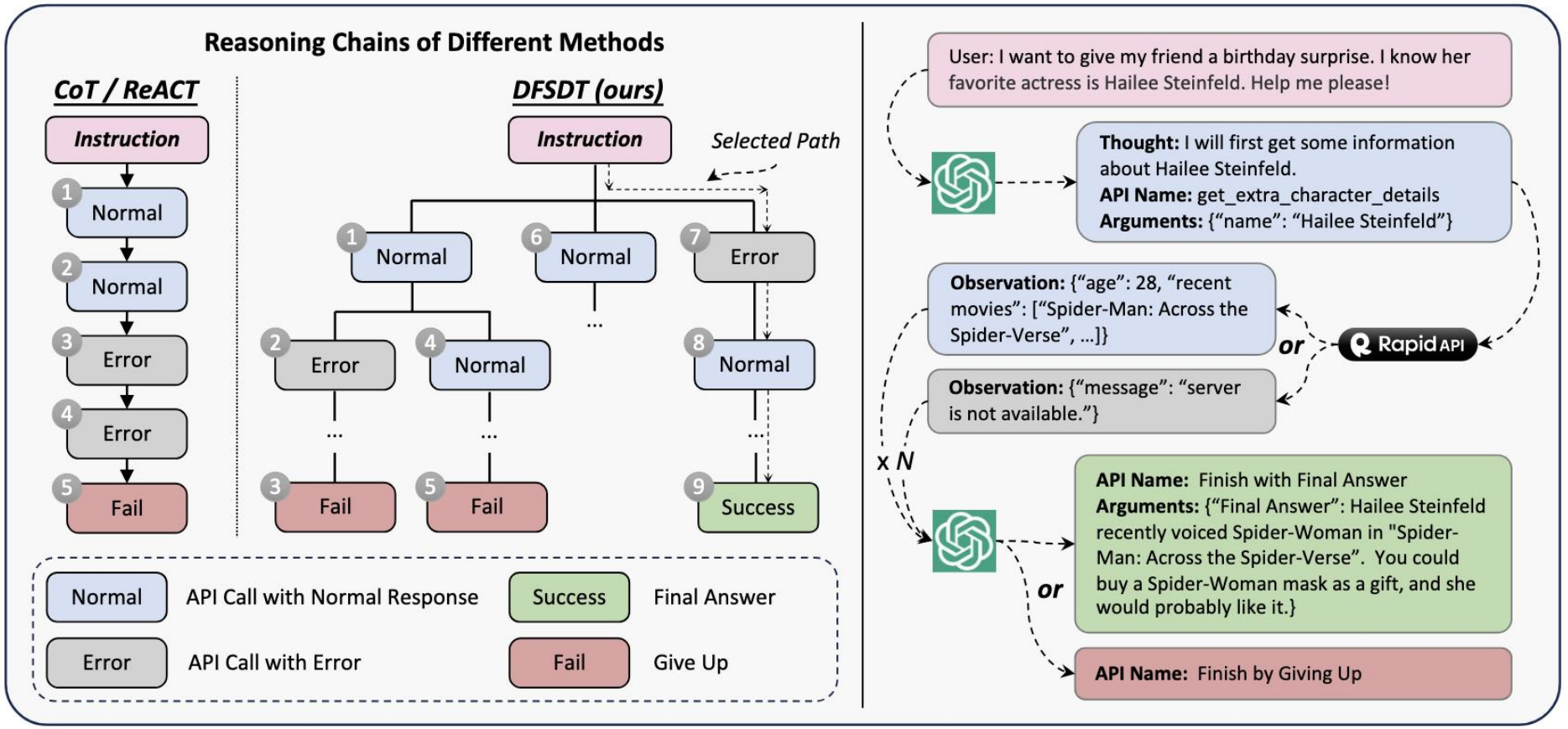
Beyond SFT: RL-Enhanced Finetuning

- MAGNET

Towards Realistic Evaluation

- TAU (τ) - Bench

Issues with ToolLLM*



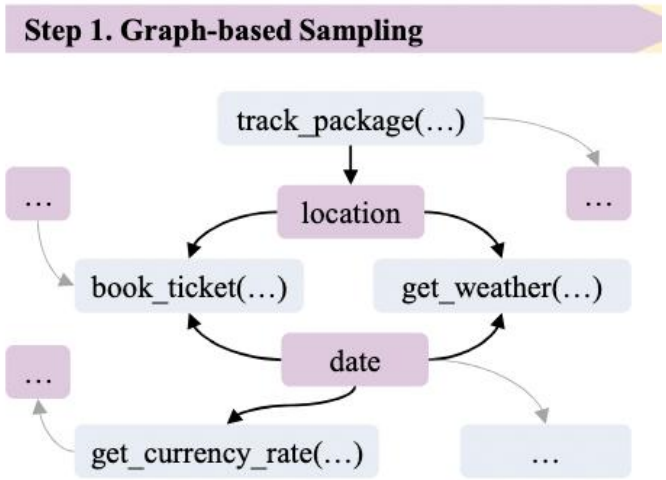
*ToolLLM: Facilitating Large Language Models to Master 16000+ Real World APIs, Qin et. al., Oct 2023

TOOLFLOW: Boosting LLM Tool-Calling Through Natural and Coherent Dialogue Synthesis

1. Framework to synthesize multi-turn function calling dialogs
2. Creates plans that guide the synthesis of coherent dialogues

* TOOLFLOW: Boosting LLM Tool-Calling Through Natural and Coherent Dialogue Synthesis, Wang et al, Mar 2025

ToolFlow: Boosting LLM Tool-Calling Through Natural and Coherent Dialogue Synthesis

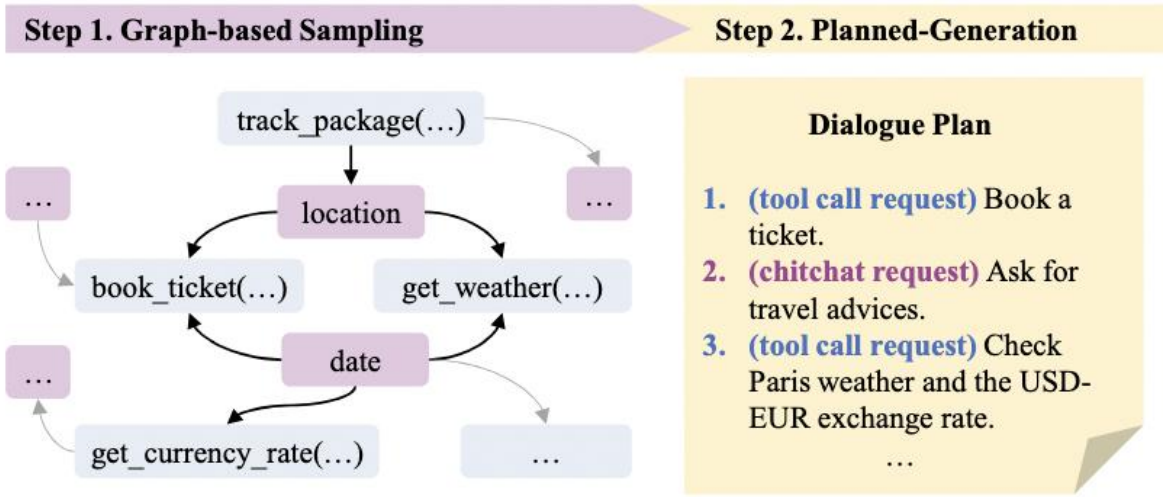


1. Parameter-Parameter Similarity
   - `location` and `destination` are semantically similar parameters, so `get_weather()` and `book_flight()` maybe used together in travel-related contexts.

2. Parameter-Return Value Similarity
   - `check_calendar()` typically returns the `location` of events, while the `navigate()` requires a `location` as input. When a user requests to *"navigate to the location of this afternoon's meeting"* both tools would be called.

* TOOLFLOW: Boosting LLM Tool-Calling Through Natural and Coherent Dialogue Synthesis, Wang et al, Mar 2025

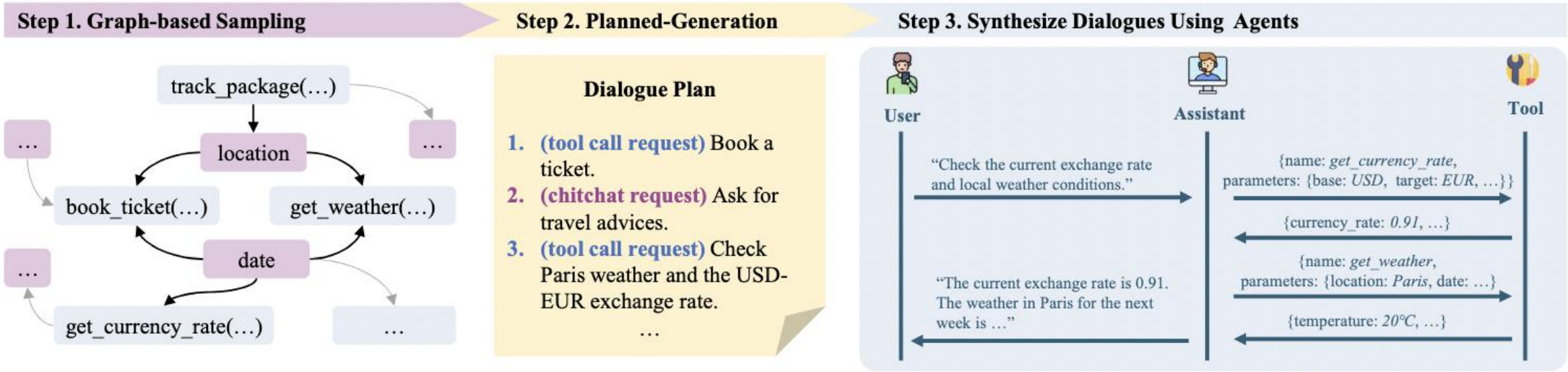
ToolFlow: Boosting LLM Tool-Calling Through Natural and Coherent Dialogue Synthesis



* TOOLFLOW: Boosting LLM Tool-Calling Through Natural and Coherent Dialogue Synthesis, Wang et al, Mar 2025

TOOLFLOW: Boosting LLM Tool-Calling Through Natural and Coherent Dialogue Synthesis

Like ToolAlpaca



* TOOLFLOW: Boosting LLM Tool-Calling Through Natural and Coherent Dialogue Synthesis, Wang et al, Mar 2025

TOOLFLOW: Boosting LLM Tool-Calling Through Natural and Coherent Dialogue Synthesis

Baselines			Proc.	Resp.	Overall
GPT-3.5†			77	85	75
ToolAlpaca-13B†			63	69	60
ToolAlpaca-7B†			70	73	70
LLaMA-3.1 8B			74	80	74
	Graph	Plan			
Ours	✓	✓	**85**	**88**	**84**
	✓	✗	80	85	80
	✗	✓	81	83	79
	✗	✗	78	83	77

* TOOLFLOW: Boosting LLM Tool-Calling Through Natural and Coherent Dialogue Synthesis, Wang et al, Mar 2025

Research Directions

High Fidelity Data Synthesis

- ToolFlow
- MAGNET

Beyond SFT: RL-Enhanced Finetuning

- MAGNET

Towards Realistic Evaluation

- TAU (τ) - Bench

Motivation

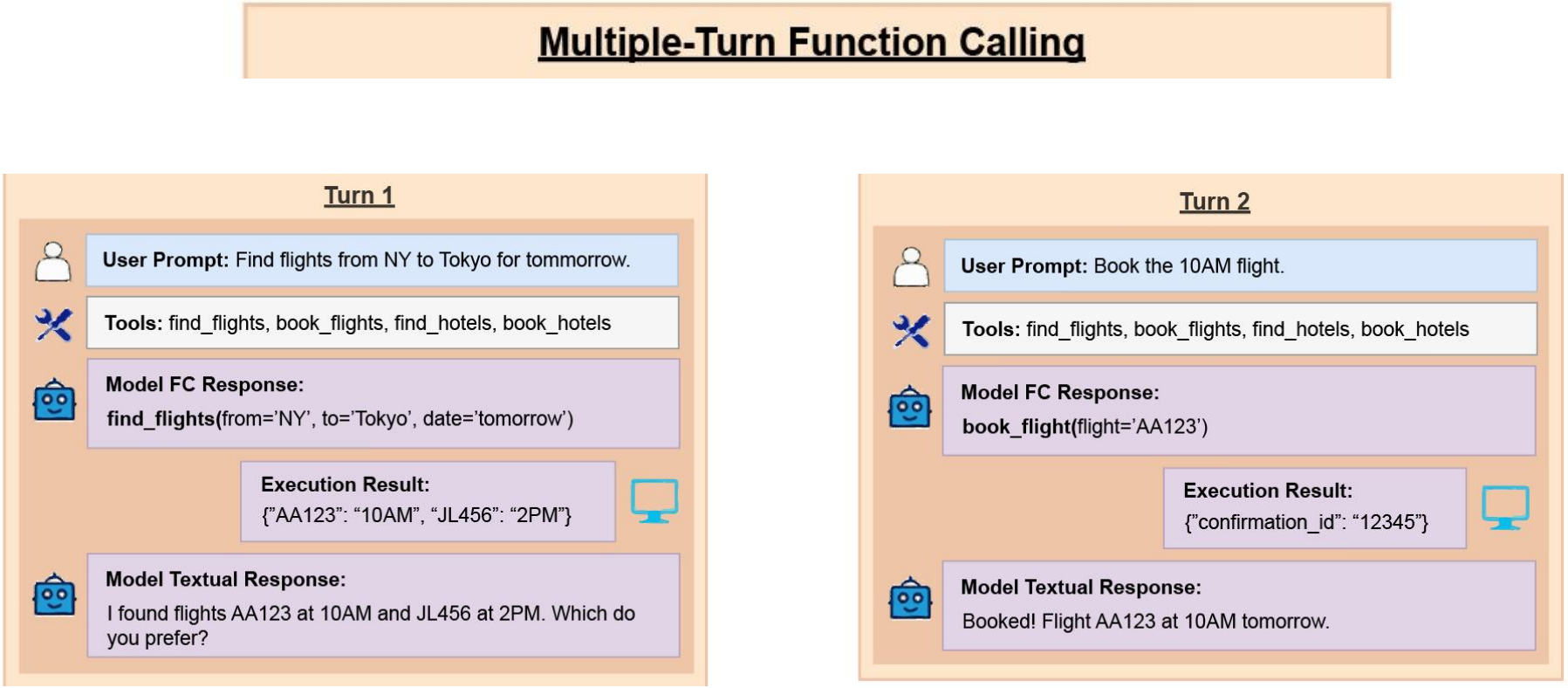


Image credits: https://gorilla.cs.berkeley.edu/blogs/13_bfcl_v3_multi_turn.html

Motivation

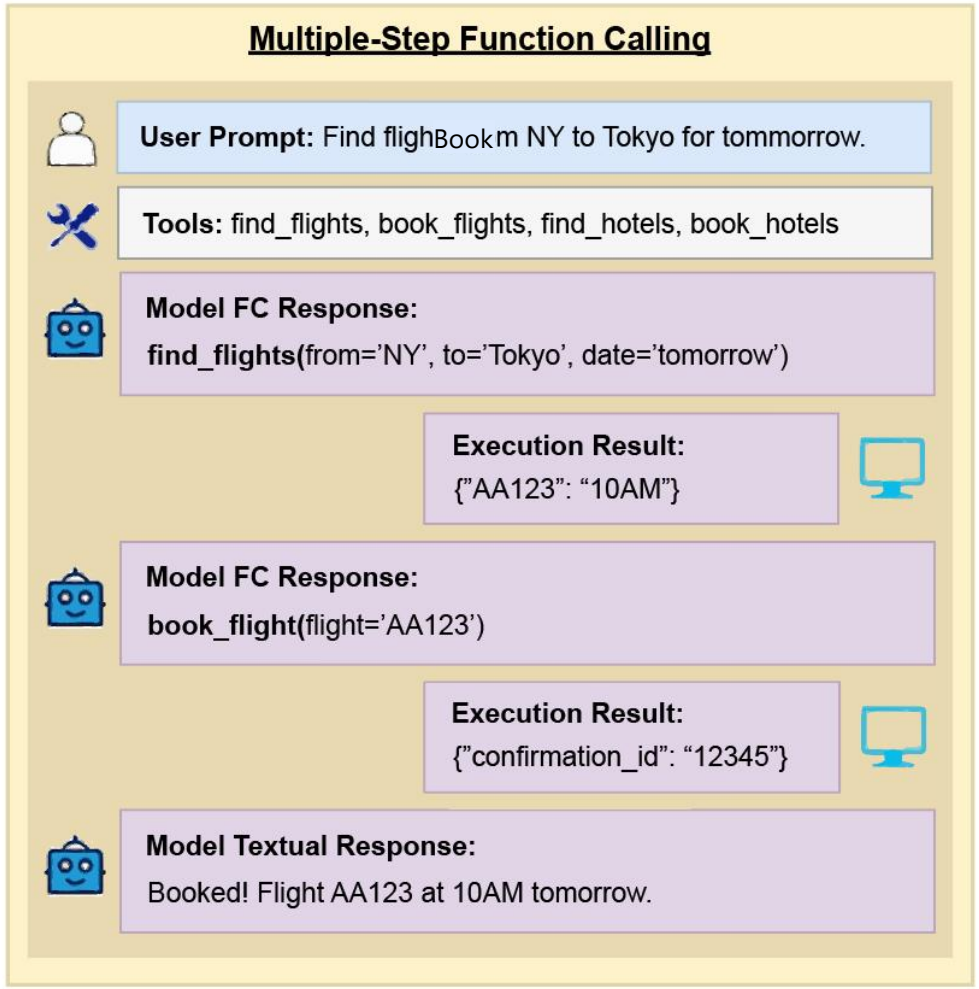


Image credits: https://gorilla.cs.berkeley.edu/blogs/13_bfcl_v3_multi_turn.html

MAGNET: Multi-turn Tool-use Data Synthesis and Distillation via Graph Translation

1. Framework to synthesize multi-turn function calling dialogs

2. Constructs training trajectories for both SFT and multi-turn DPO (mDPO)

* Magnet: Multi-turn Tool-use Data Synthesis and Distillation via Graph Translation, Yin et al, Mar 2025

MAGNET: Multi-turn Tool-use Data Synthesis and Distillation via Graph Translation*

Let

- H be the total number of turns in a dialog
- u_h be the user input at turn h
- a_h be the model action at turn h
- t_h be the tool response at turn h

A trajectory (τ) involves a sequence (H -turns) of user inputs, model actions, and tool response

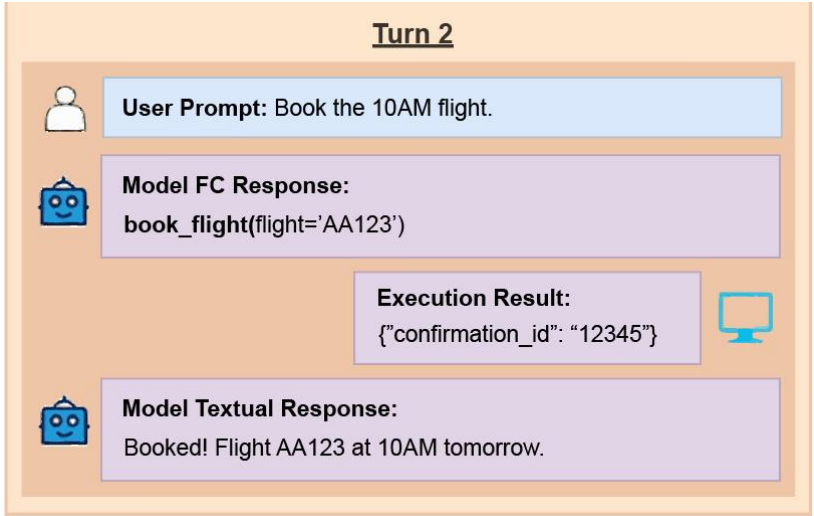
$$\tau = (u_1, a_1, t_1, \cdots, u_{\mathrm{H}}, a_{\mathrm{H}}, t_{\mathrm{H}})$$

* Magnet: Multi-turn Tool-use Data Synthesis and Distillation via Graph Translation, Yin et al, Mar 2025

MAGNET: Multi-turn Tool-use Data Synthesis and Distillation via Graph Translation*

Let

- H be the total number of turns in a dialog
- u_h be the user input at turn h - **some user inputs can be skipped**
- a_h be the model action at turn h
- t_h be the tool response at turn h



A trajectory (τ) involves a sequence (H -turns) of user inputs, model actions, and tool response

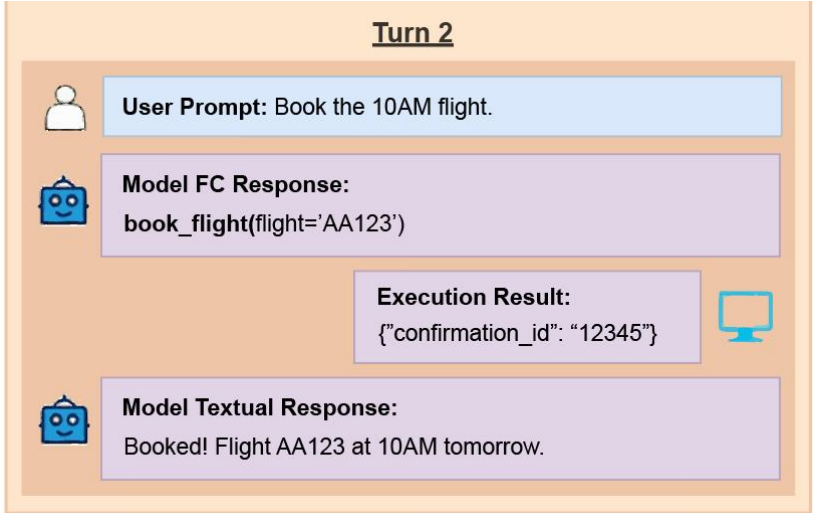
$$\tau = (u_1, a_1, t_1, \cdots, u_H, a_H, t_H)$$

* Magnet: Multi-turn Tool-use Data Synthesis and Distillation via Graph Translation, Yin et al, Mar 2025

MAGNET: Multi-turn Tool-use Data Synthesis and Distillation via Graph Translation*

Let

- H be the total number of turns in a dialog
- u_h be the user input at turn h - some user inputs can be skipped
- a_h be the model action at turn h - **actions can be NL response or too**
- t_h be the tool response at turn h



A trajectory (τ) involves a sequence (H -turns) of user inputs, model actions, and tool response

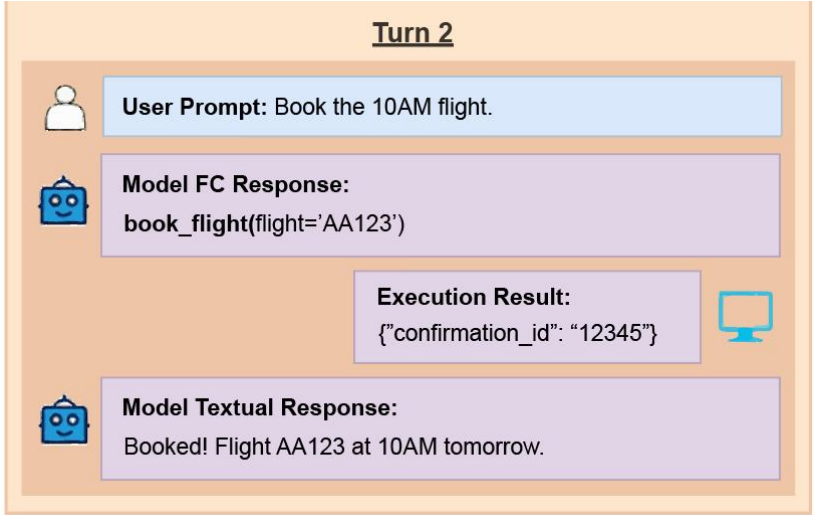
$$\tau = (u_1, a_1, t_1, \cdots, u_H, a_H, t_H)$$

* Magnet: Multi-turn Tool-use Data Synthesis and Distillation via Graph Translation, Yin et al, Mar 2025

MAGNET: Multi-turn Tool-use Data Synthesis and Distillation via Graph Translation*

Let

- H be the total number of turns in a dialog
- u_h be the user input at turn h - some user inputs can be skipped
- a_h be the model action at turn h - actions can be NL response or too
- t_h be the tool response at turn h
  - **tool response following NL action is NULL**



A trajectory (τ) involves a sequence (H -turns) of user inputs, model actions, and tool response

$$\tau = (u_1, a_1, t_1, \cdots, u_H, a_H, t_H)$$

* Magnet: Multi-turn Tool-use Data Synthesis and Distillation via Graph Translation, Yin et al, Mar 2025

MAGNET: Multi-turn Tool-use Data Synthesis and Distillation via Graph Translation*

Positive Trajectory

$$\tau_w = (u_1^w, a_1^w, t_1^w, \cdots, u_{H_w}^w, a_{H_w}^w, t_{H_w}^w)$$

Negative Trajectory

$$\tau_l = (u_1^l, a_1^l, t_1^l, \cdots, u_{H_l}^l, a_{H_l}^l, t_{H_l}^l)$$

* Magnet: Multi-turn Tool-use Data Synthesis and Distillation via Graph Translation, Yin et al, Mar 2025

MAGNET: Multi-turn Tool-use Data Synthesis and Distillation via Graph Translation*

Positive Trajectory

$$\tau_w = (u_1^w, a_1^w, t_1^w, \cdots, u_{H_w}^w, a_{H_w}^w, t_{H_w}^w)$$

Negative Trajectory

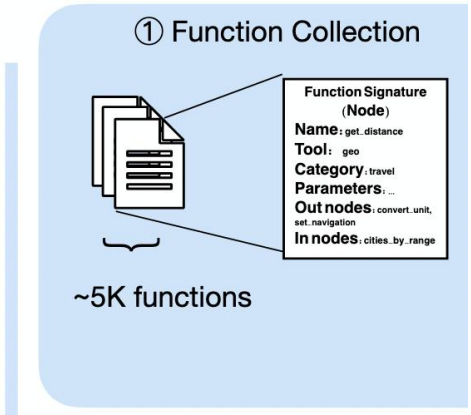
$$\tau_l = (u_1^l, a_1^l, t_1^l, \cdots, u_{H_l}^l, a_{H_l}^l, t_{H_l}^l)$$

$$\mathcal{L}(x; \tau_w, \tau_l) = \mathcal{L}_{\text{SFT}}(x; \tau_w) + \lambda \mathcal{L}_{\text{mDPO}}(x; \tau_w, \tau_l),$$

$$\mathcal{L}_{\text{mDPO}}(x; \tau_w, \tau_l) = -\log\sigma\left(\eta\left(\sum_{\tau_l} \frac{\pi_\theta(a^l|s^l)}{\pi_{ref}(a^l|s^l)} - \sum_{\tau_w} \frac{\pi_\theta(a^w|s^w)}{\pi_{ref}(a^w|s^w)}\right)\right)$$

* Magnet: Multi-turn Tool-use Data Synthesis and Distillation via Graph Translation, Yin et al, Mar 2025

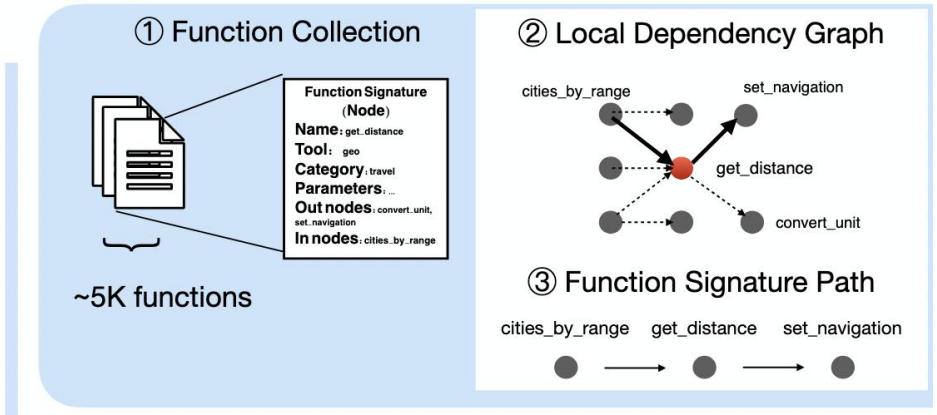
MAGNET: Multi-turn Tool-use Data Synthesis and Distillation via Graph Translation*



1. StableToolBench (Guo et al., 2024) - RapidAPIs with Cache
2. BFCL-v3 multi-turn function implementation (Yan et al., 2024)

* Magnet: Multi-turn Tool-use Data Synthesis and Distillation via Graph Translation, Yin et al, Mar 2025

MAGNET: Multi-turn Tool-use Data Synthesis and Distillation via Graph Translation*

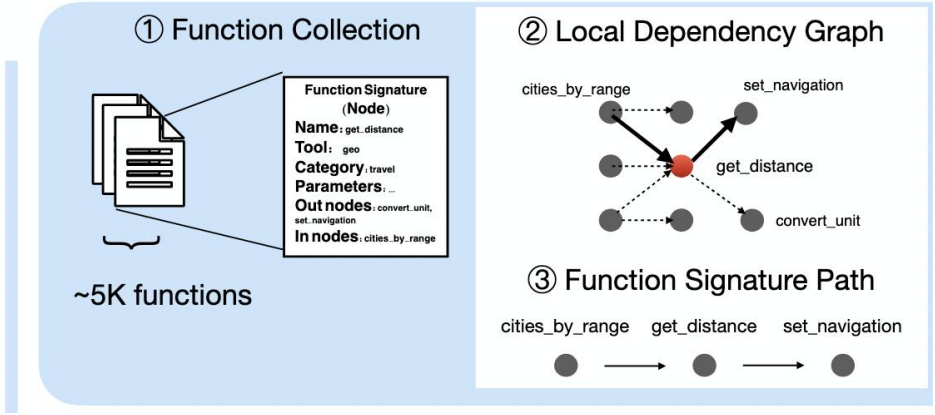


Graph Construction:

- 5k functions are nodes
- Assign labels to each tool using LLM
- For a given node, randomly sample nodes with same label as candidate neighbors
- Use LLM as a judge to identify good candidate neighbors

* Magnet: Multi-turn Tool-use Data Synthesis and Distillation via Graph Translation, Yin et al, Mar 2025

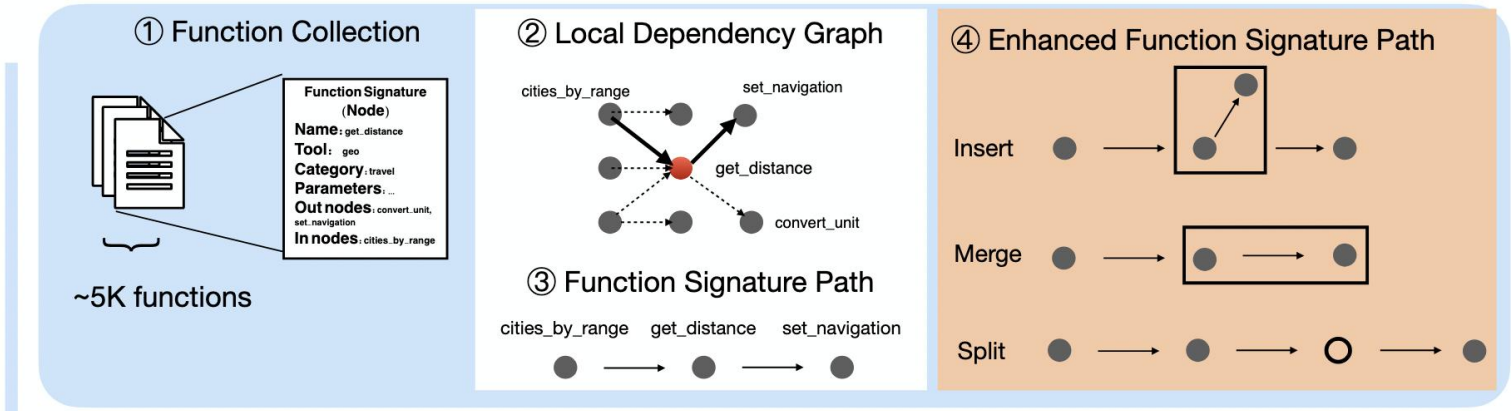
MAGNET: Multi-turn Tool-use Data Synthesis and Distillation via Graph Translation*



$$\mathrm{FSP} = (\mathrm{fs}_1, \mathrm{fs}_2, \cdots, \mathrm{fs}_\mathrm{H})$$

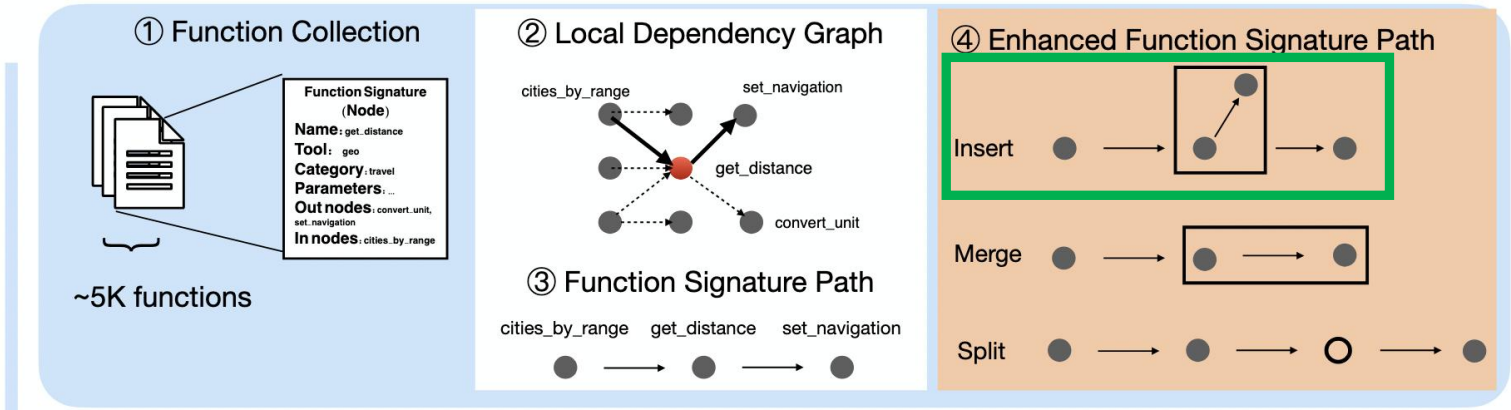
* Magnet: Multi-turn Tool-use Data Synthesis and Distillation via Graph Translation, Yin et al, Mar 2025

MAGNET: Multi-turn Tool-use Data Synthesis and Distillation via Graph Translation*



* Magnet: Multi-turn Tool-use Data Synthesis and Distillation via Graph Translation, Yin et al, Mar 2025

MAGNET: Multi-turn Tool-use Data Synthesis and Distillation via Graph Translation*

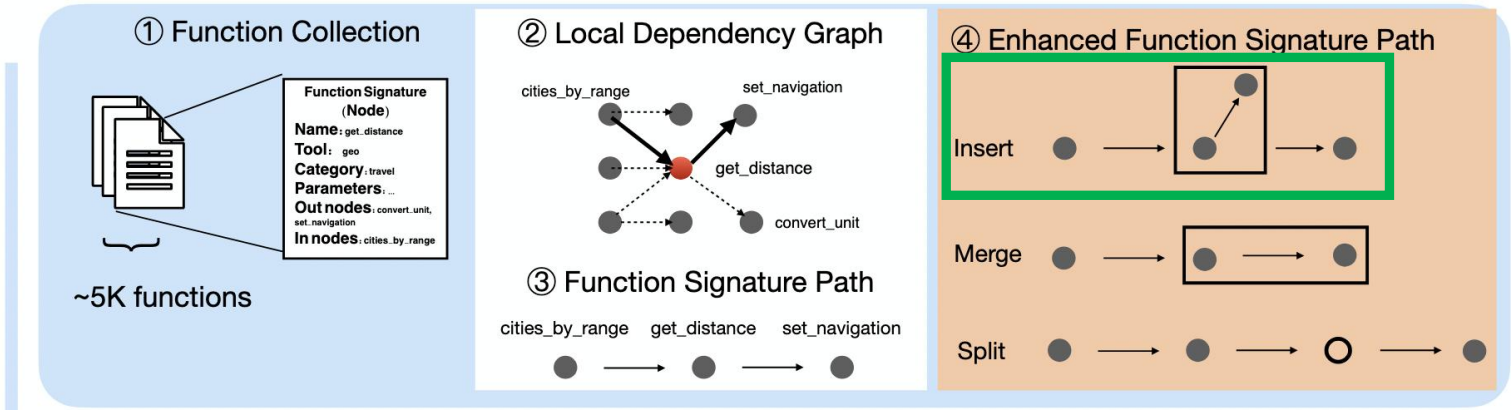


$$\text{FSP} = (\text{fs}_1, \text{fs}_2, \cdots, \text{fs}_\text{H})$$

$$\text{FSP} = (\text{fs}_1, \{\text{fs}_{21}, \text{fs}_{22}\}, \cdots, \text{fs}_\text{H})$$

* Magnet: Multi-turn Tool-use Data Synthesis and Distillation via Graph Translation, Yin et al, Mar 2025

MAGNET: Multi-turn Tool-use Data Synthesis and Distillation via Graph Translation*

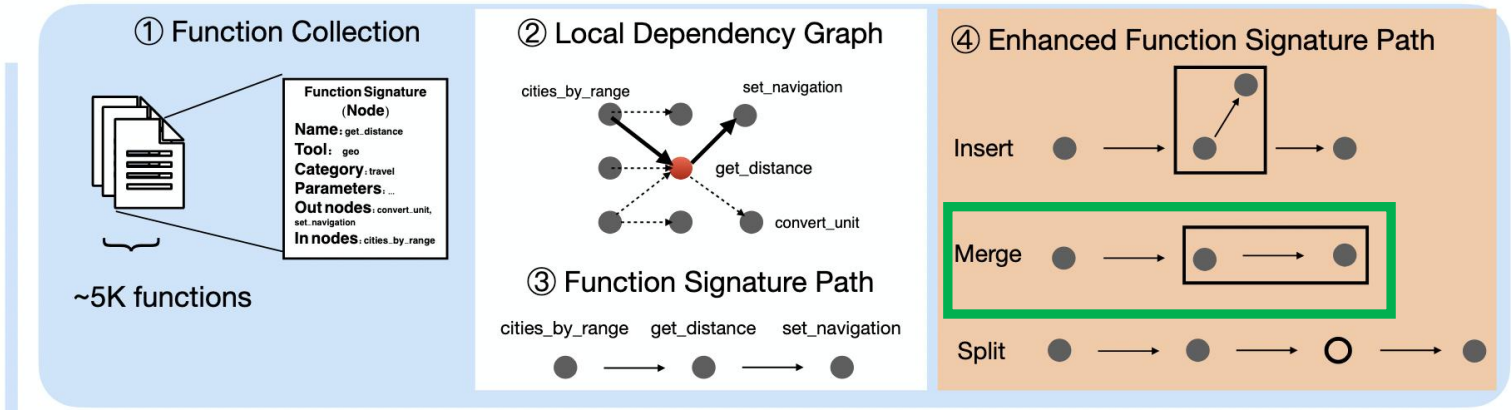



```
fs2: get_flight_by_airport(airport_symbol)
```

```
fs2: {get_airport_symbol_by_city(city,range_in_kms), get_flight_by_airport(airport_symbol)}
```

* Magnet: Multi-turn Tool-use Data Synthesis and Distillation via Graph Translation, Yin et al, Mar 2025

MAGNET: Multi-turn Tool-use Data Synthesis and Distillation via Graph Translation*

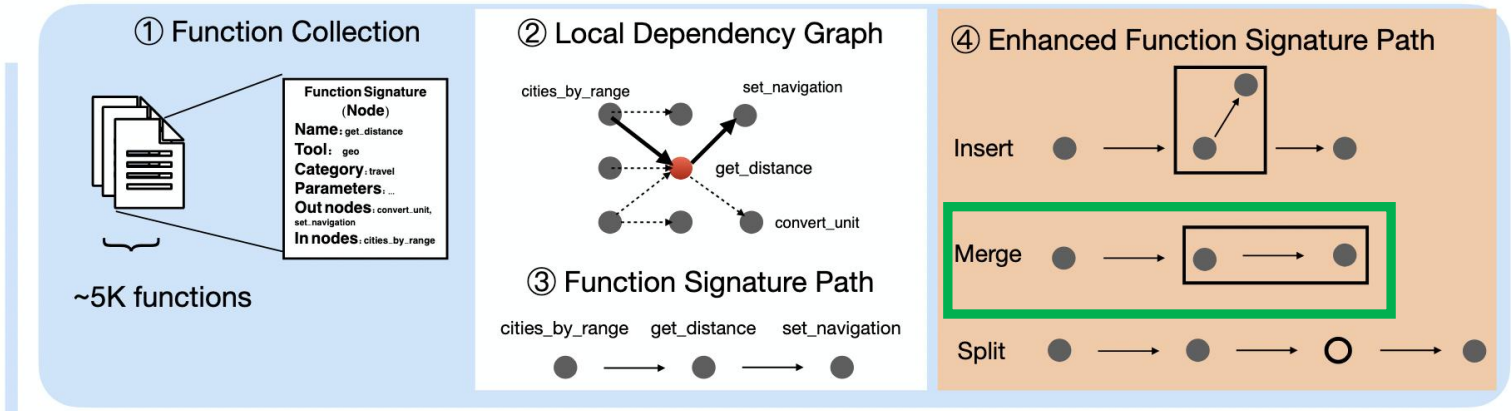


$$\text{FSP} = (\text{fs}_1, \text{fs}_2, \text{fs}_3, \cdots, \text{fs}_\text{H})$$

$$\text{FSP} = (\{\text{fs}_{11}, \text{fs}_{12}\}, \text{fs}_2, \cdots, \text{fs}_{\text{H}-1})$$

* Magnet: Multi-turn Tool-use Data Synthesis and Distillation via Graph Translation, Yin et al, Mar 2025

MAGNET: Multi-turn Tool-use Data Synthesis and Distillation via Graph Translation*



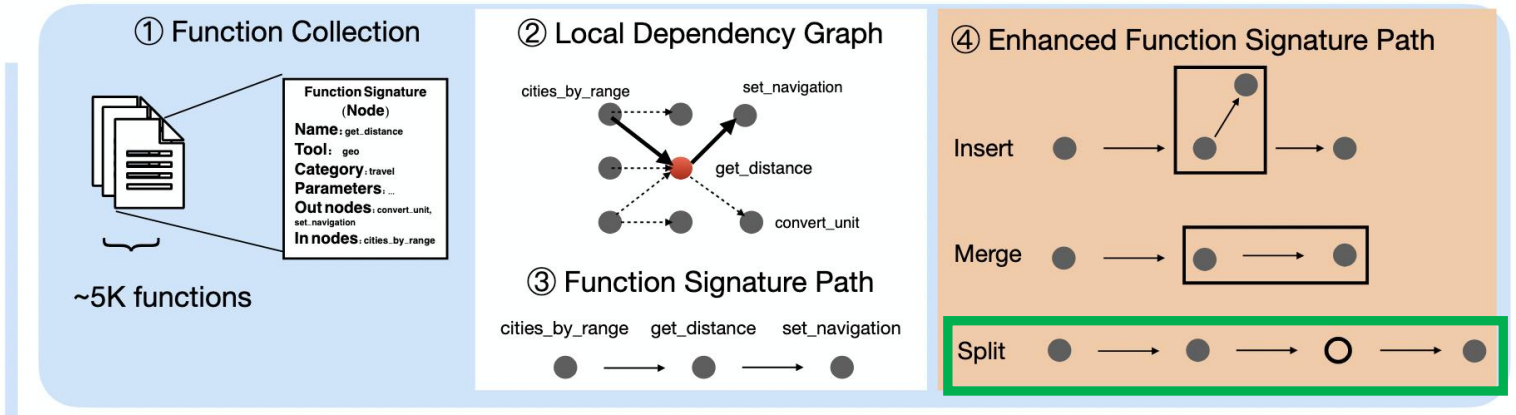
fs2: `get_distance(from_loc,to_loc)`

fs3: `set_navigation(distance)`

fs2: `{get_distance(from_loc,to_loc),set_navigation(distance)}`

* Magnet: Multi-turn Tool-use Data Synthesis and Distillation via Graph Translation, Yin et al, Mar 2025

MAGNET: Multi-turn Tool-use Data Synthesis and Distillation via Graph Translation*

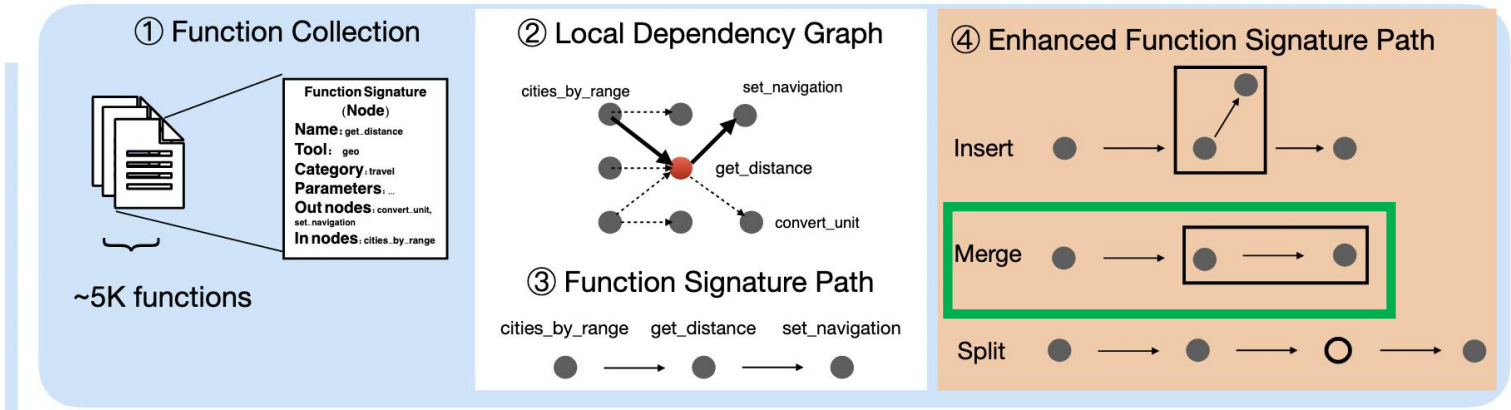


$$\text{FSP} = (\text{fs}_1, \text{fs}_2, \cdots, \text{fs}_\text{H})$$

$$\text{FSP} = (\text{fs}_1, \text{fs}_{21}, \text{fs}_{22}, \cdots, \text{fs}_\text{H})$$

* Magnet: Multi-turn Tool-use Data Synthesis and Distillation via Graph Translation, Yin et al, Mar 2025

MAGNET: Multi-turn Tool-use Data Synthesis and Distillation via Graph Translation*



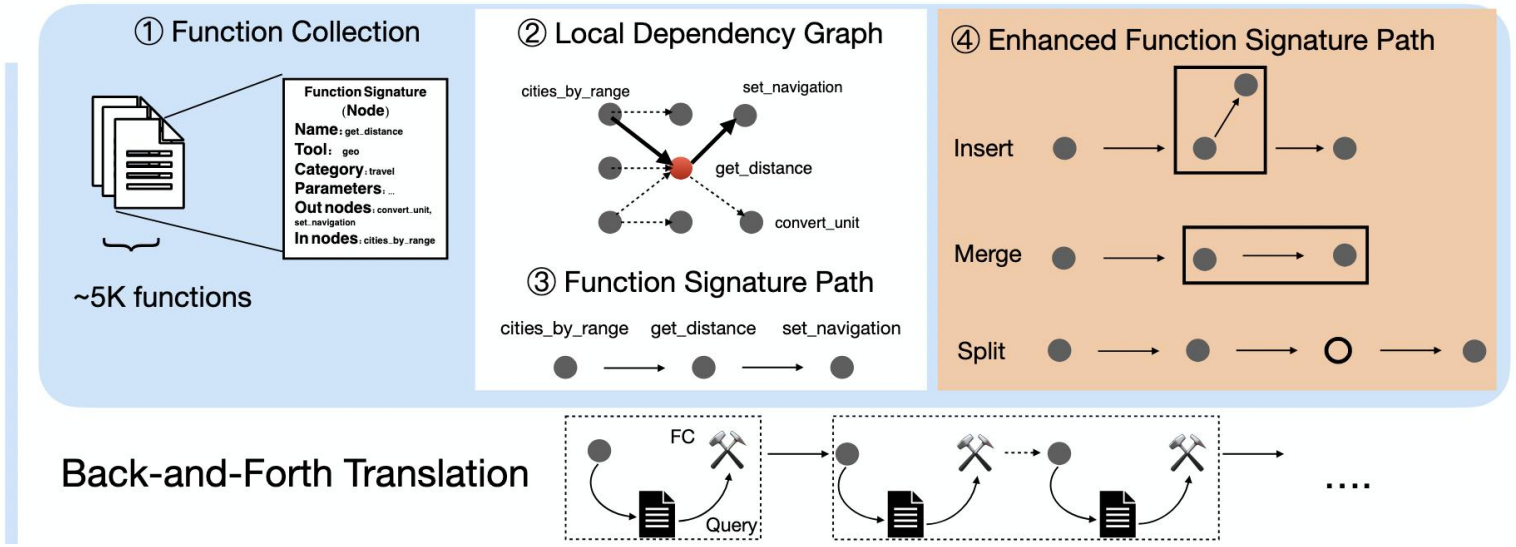
fs2: `get_distance(from_loc,to_loc)`

fs21: `get_distance(?,to_loc)` without `from_loc`

fs22: received `from_loc`

* Magnet: Multi-turn Tool-use Data Synthesis and Distillation via Graph Translation, Yin et al, Mar 2025

MAGNET: Multi-turn Tool-use Data Synthesis and Distillation via Graph Translation*

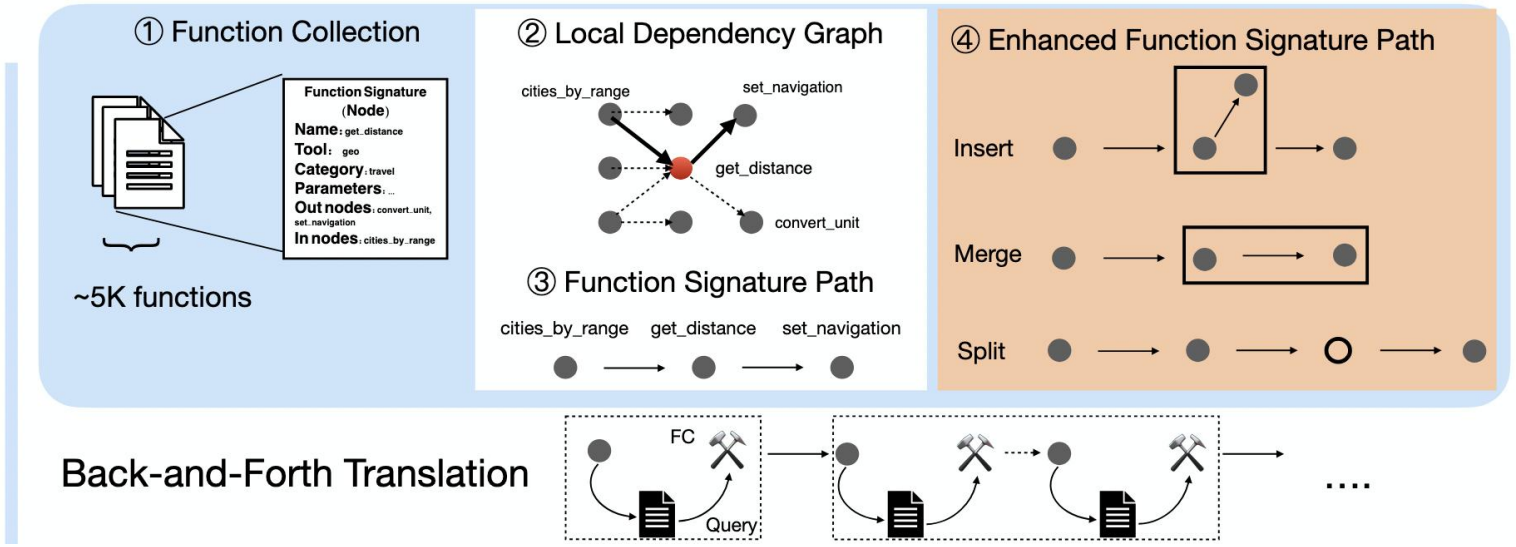


$$\text{FSP} = (\text{fs}_1, \text{fs}_2, \cdots, \text{fs}_\text{H})$$

$$\text{Back Translation} : \mathcal{M}_\text{b}(\text{fs}_\text{h}) = u_h$$

* Magnet: Multi-turn Tool-use Data Synthesis and Distillation via Graph Translation, Yin et al, Mar 2025

MAGNET: Multi-turn Tool-use Data Synthesis and Distillation via Graph Translation*



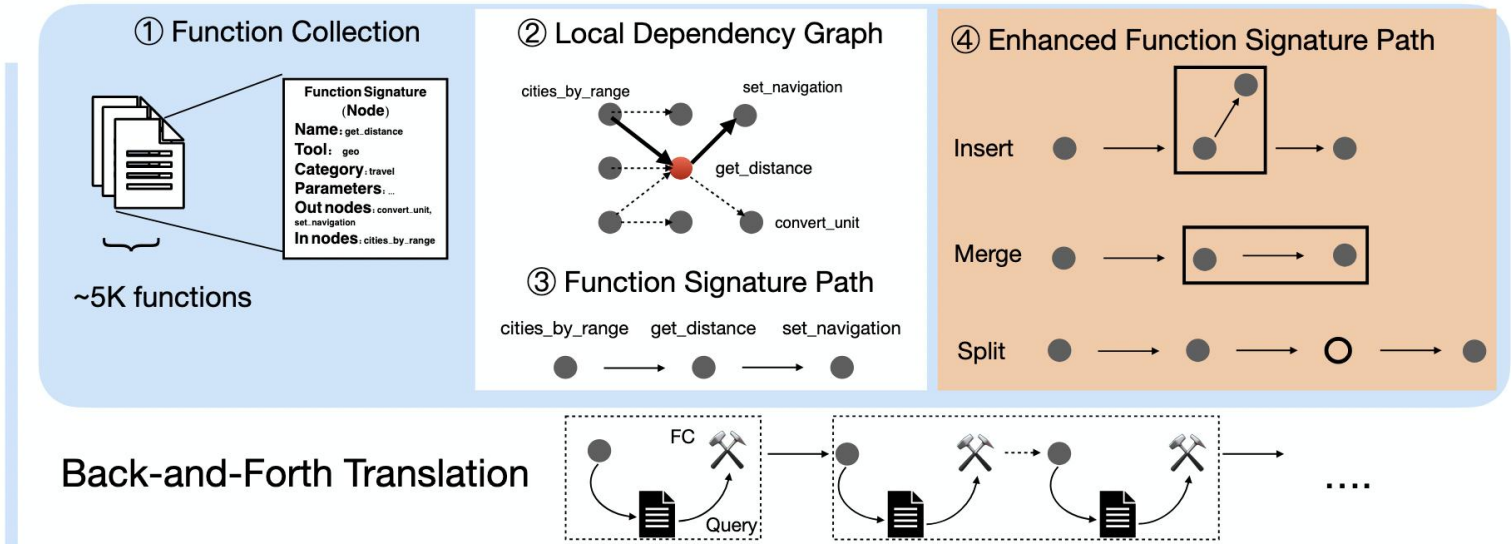
$$\text{FSP} = (\text{fs}_1, \text{fs}_2, \cdots, \text{fs}_\text{H})$$

$$\tau = (u_1, a_1, t_1, \cdots, u_\text{H}, a_\text{H}, t_\text{H})$$

Back Translation : $\mathcal{M}_\text{b}(\text{fs}_\text{h}) = u_h$

* Magnet: Multi-turn Tool-use Data Synthesis and Distillation via Graph Translation, Yin et al, Mar 2025

MAGNET: Multi-turn Tool-use Data Synthesis and Distillation via Graph Translation*

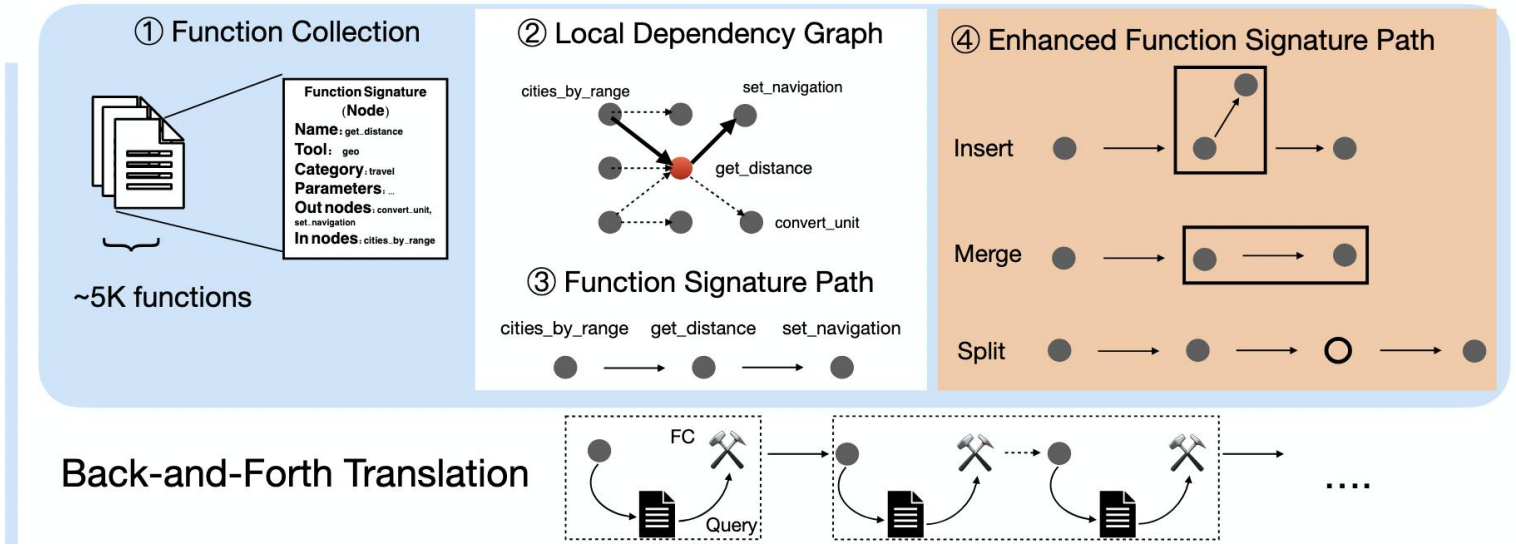



```
fs1: get_flight_by_airport(airport_symbol)
```

Back Translation $\mathcal{M}_b(fs_1)$ = "What all flights are landing in New Delhi (DEL)?"

* Magnet: Multi-turn Tool-use Data Synthesis and Distillation via Graph Translation, Yin et al, Mar 2025

MAGNET: Multi-turn Tool-use Data Synthesis and Distillation via Graph Translation*



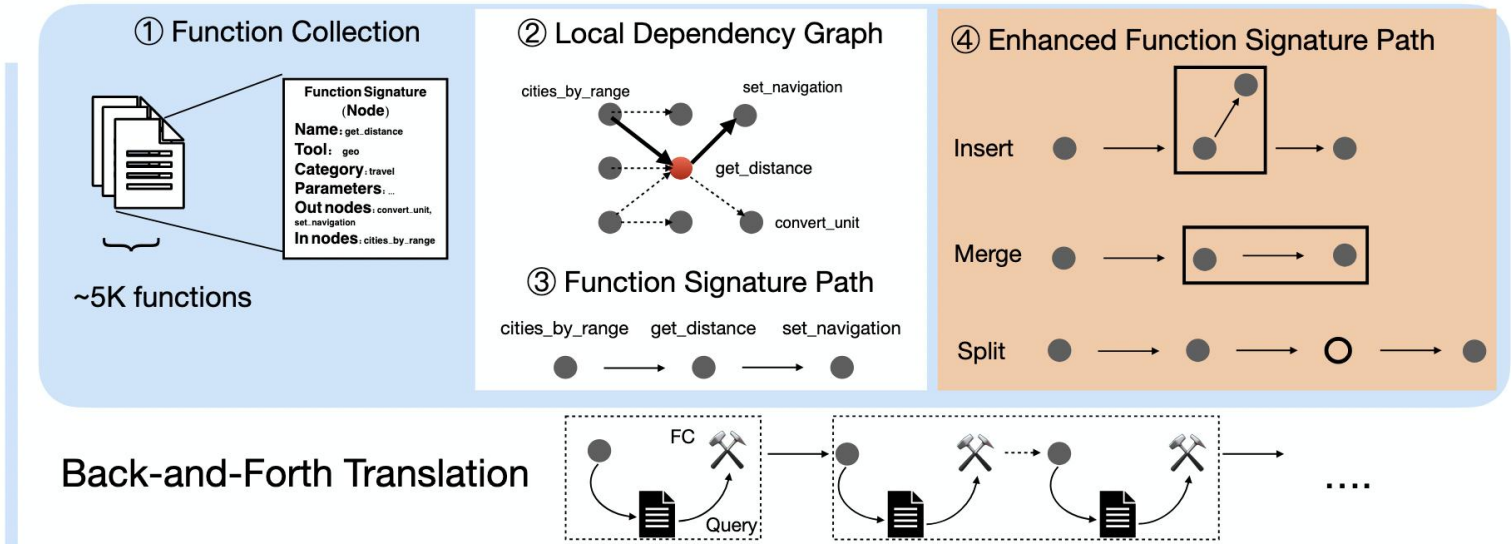
$$\mathrm{FSP} = (\mathrm{fs}_1, \mathrm{fs}_2, \cdots, \mathrm{fs}_\mathrm{H})$$

$$\tau = (u_1, a_1, t_1, \cdots, u_\mathrm{H}, a_\mathrm{H}, t_\mathrm{H})$$

Forth Translation : $\mathcal{M}_\mathrm{b}(\mathrm{t}_{\mathrm{h}-1}, \mathrm{u}_\mathrm{h}, \mathrm{fs}_\mathrm{h}) = \mathrm{f}_\mathrm{h}$

* Magnet: Multi-turn Tool-use Data Synthesis and Distillation via Graph Translation, Yin et al, Mar 2025

MAGNET: Multi-turn Tool-use Data Synthesis and Distillation via Graph Translation*



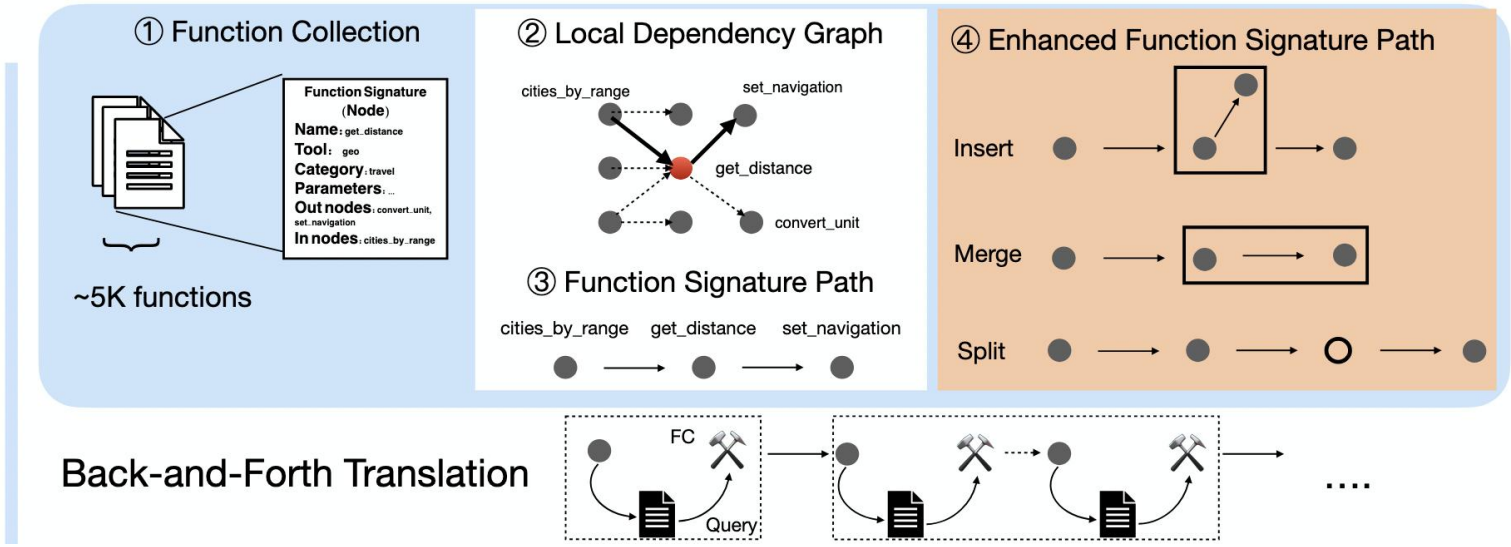
fs1: `get_flight_by_airport(airport_symbol)`

u_1: "What all flights are landing in New Delhi (DEL)?"

Forth Translation : $\mathcal{M}_b(t_0, u_1, fs_1)$ = `get_flight_by_airport(airport_symbol=DEL)`

* Magnet: Multi-turn Tool-use Data Synthesis and Distillation via Graph Translation, Yin et al, Mar 2025

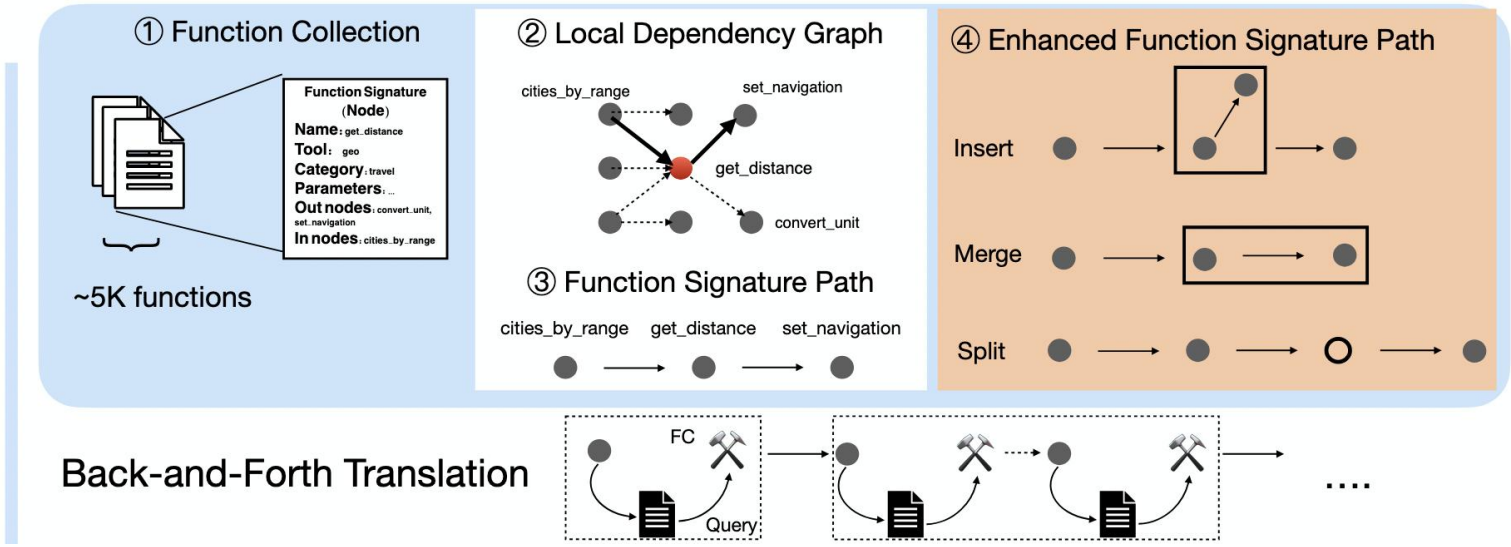
MAGNET: Multi-turn Tool-use Data Synthesis and Distillation via Graph Translation*



$$\tau = (u_1, a_1, t_1, \cdots, u_{\mathrm{H}}, a_{\mathrm{H}}, t_{\mathrm{H}})$$

* Magnet: Multi-turn Tool-use Data Synthesis and Distillation via Graph Translation, Yin et al, Mar 2025

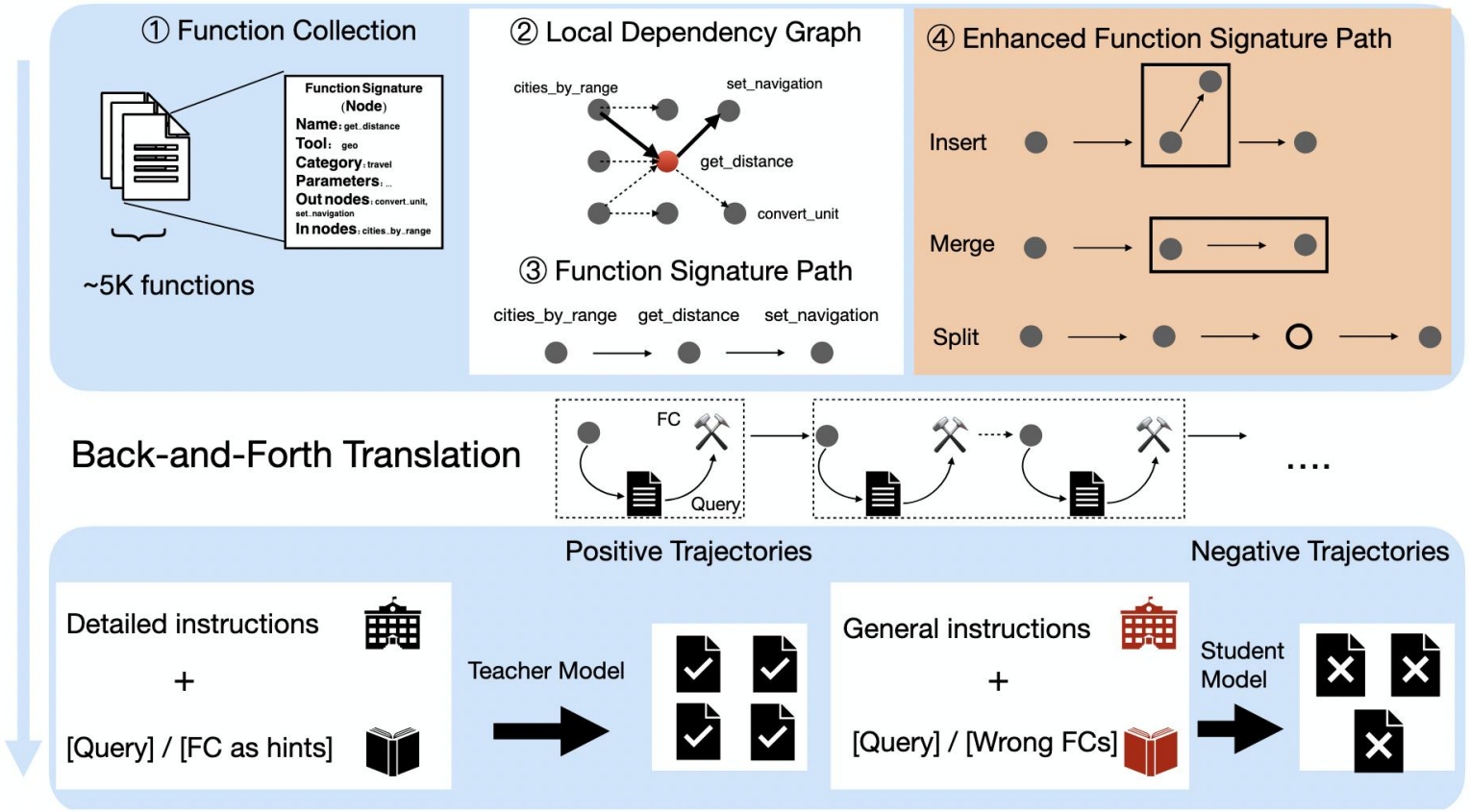
MAGNET: Multi-turn Tool-use Data Synthesis and Distillation via Graph Translation*



$$\tau = (u_1, a_1, t_1, \cdots, u_{\mathrm{H}}, a_{\mathrm{H}}, t_{\mathrm{H}})$$

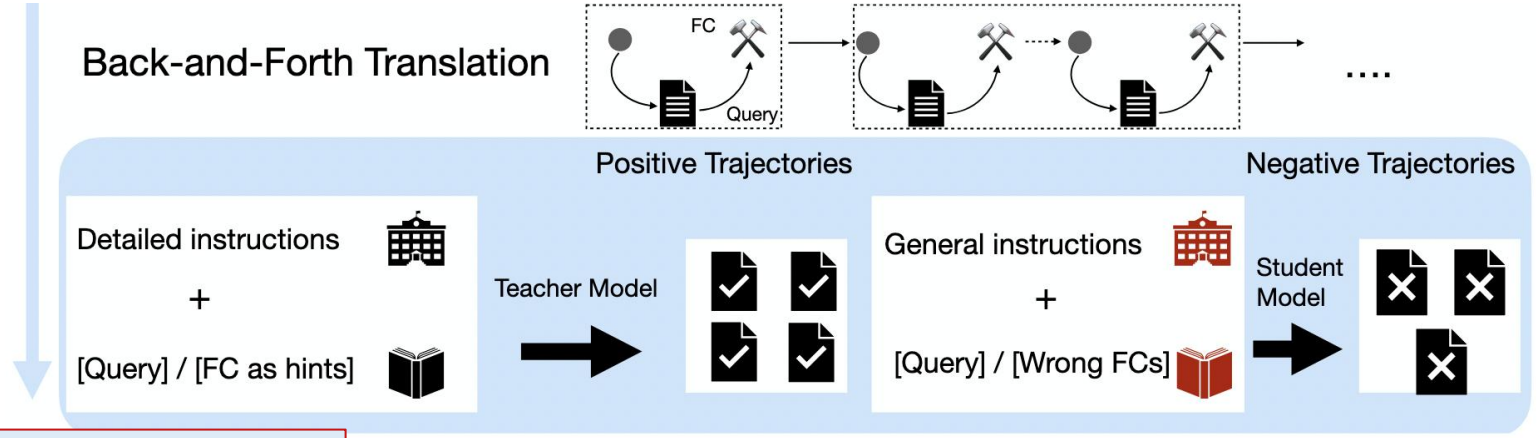
* Magnet: Multi-turn Tool-use Data Synthesis and Distillation via Graph Translation, Yin et al, Mar 2025

MAGNET: Multi-turn Tool-use Data Synthesis and Distillation via Graph Translation*



* Magnet: Multi-turn Tool-use Data Synthesis and Distillation via Graph Translation, Yin et al, Mar 2025

MAGNET: Multi-turn Tool-use Data Synthesis and Distillation via Graph Translation*



$$\tau = (u_1, a_1, t_1, \cdots, u_H, a_H, t_H)$$

fs1: `get_flight_by_airport(airport_symbol)`

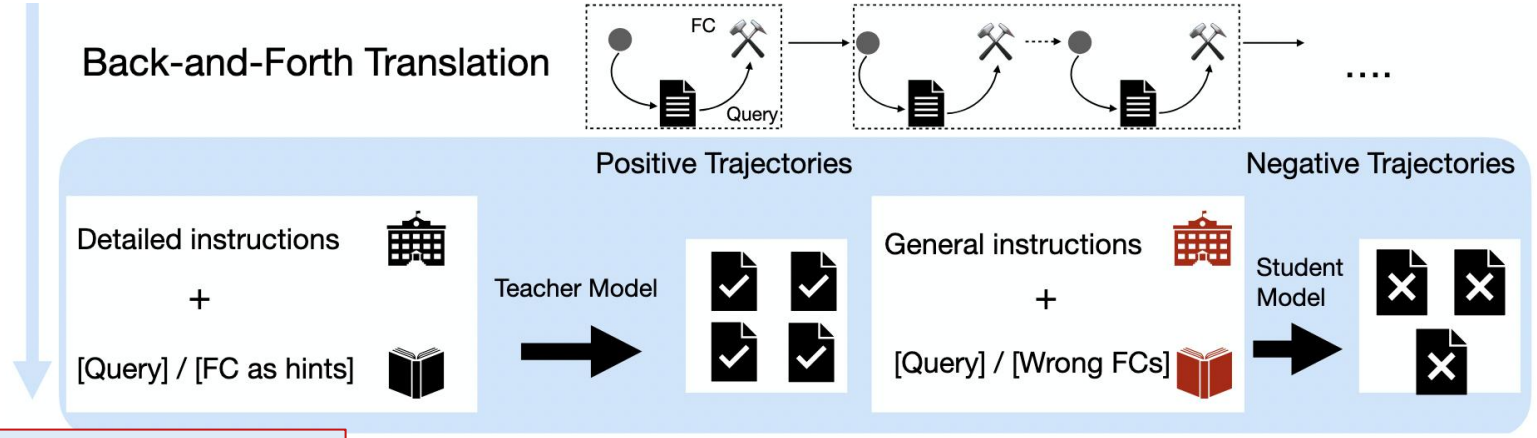
u_1: "What all flights are landing in New Delhi (DEL)?"

Forth Translation : $\mathcal{M}_b(t_0, u_1, fs_1)$ = `get_flight_by_airport(airport_symbol=DEL)`

a_1: `get_flight_by_airport(airport_symbol=DEL)`

* Magnet: Multi-turn Tool-use Data Synthesis and Distillation via Graph Translation, Yin et al, Mar 2025

MAGNET: Multi-turn Tool-use Data Synthesis and Distillation via Graph Translation*



$\tau = (u_1, a_1, t_1, \cdots, u_H, a_H, t_H)$

fs1: `get_distance(?,to_loc)` without `from_loc`

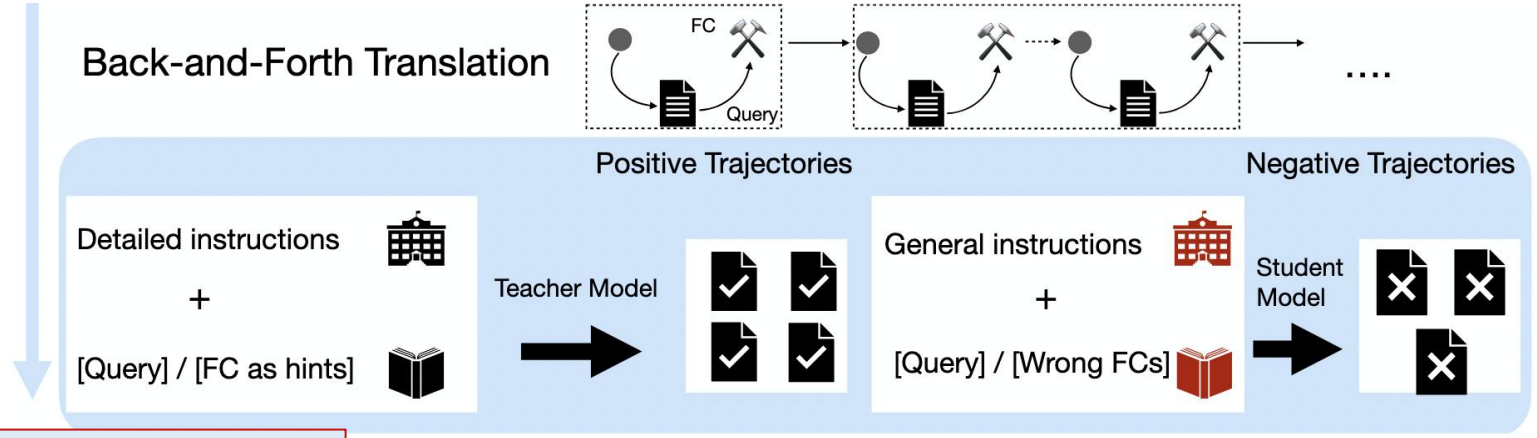
u_1: "How far is Agra?"

Forth Translation : $\mathcal{M}_b(t_0, u_1, fs_1)$ = `get_distance(from_loc=?, to_loc=agra)`

a_1: Could you tell me where you're starting from?

* Magnet: Multi-turn Tool-use Data Synthesis and Distillation via Graph Translation, Yin et al, Mar 2025

MAGNET: Multi-turn Tool-use Data Synthesis and Distillation via Graph Translation*



$\tau = (u_1, a_1, t_1, \cdots, u_H, a_H, t_H)$

fs1: `get_distance(?,to_loc)` without `from_loc`

u_1: "How far is Agra?"

Forth Translation : $\mathcal{M}_b(t_0, u_1, fs_1) =$ `get_weather(location=agra)`

a_1: `get_weather(location=agra)`

* Magnet: Multi-turn Tool-use Data Synthesis and Distillation via Graph Translation, Yin et al, Mar 2025

MAGNET: Multi-turn Tool-use Data Synthesis and Distillation via Graph Translation*

Model	Overall	Single Turn			Multi-turn				
		Non-live AST	Non-live Exec	Live AST	Overall	Base	Miss Func	Miss Param	Long
Top six models									
watt-tool-70B (FC)	74.31	84.06	89.39	77.74	58.75	67.50	57.50	48.50	61.50
gpt-4o-2024-11-20 (Prompt)	72.08	88.10	89.38	79.83	47.62	59.00	41.00	35.50	55.00
gpt-4o-2024-11-20 (FC)	69.58	87.42	89.20	79.65	41.00	62.50	6.00	37.50	58.00
GPT-4-turbo-2024-04-09	67.88	84.73	85.21	80.50	38.12	54.00	13.50	35.50	49.50
watt-tool-8B* (FC)	67.33	86.44	87.73	76.23	38.25	46.00	40.00	27.00	40.00
o1-2024-12-17 (Prompt)	66.73	78.92	82.70	78.14	28.25	40.50	5.00	34.50	33.00
Gemini models (teachers)									
Gemini-1.5-Pro-002 (Prompt)	62.19	88.58	91.27	76.72	20.75	23.00	19.50	17.50	23.00
Gemini-2.0-Flash-Exp (Prompt)	61.74	89.96	79.89	82.01	17.88	28.00	3.00	19.00	21.50
7B models									
Functionary-Small-v3.1 (FC)	56.49	86.75	87.12	73.75	10.12	18.00	2.50	14.00	6.00
Hammer2.1-7b (FC)	61.83	88.65	85.48	75.11	23.50	35.50	**25.50**	19.00	14.00
Qwen2.5-Coder-7B-Instruct	53.13	86.83	82.27	66.99	8.25	11.50	6.50	5.50	5.50
Magnet-7B-SFT	62.73	88.60	85.73	74.19	26.50	35.50	24.00	**27.50**	19.00
Magnet-7B-mDPO	**64.64**	**89.40**	**89.27**	**77.92**	**27.75**	**39.00**	24.00	26.00	**22.00**
14B models									
Qwen2.5-Coder-14B-Instruct	51.88	**90.94**	87.80	65.30	5.38	7.50	7.00	4.00	3.00
MAGNET-14B-SFT	66.83	90.02	88.20	77.92	33.38	47.00	32.00	32.00	22.50
Magnet-14B-mDPO	**68.01**	90.13	**89.75**	**79.14**	**37.88**	**52.00**	**36.00**	**35.50**	**28.00**

* Magnet: Multi-turn Tool-use Data Synthesis and Distillation via Graph Translation, Yin et al, Mar 2025

Research Directions

High Fidelity Data Synthesis

- ToolFlow
- MAGNET

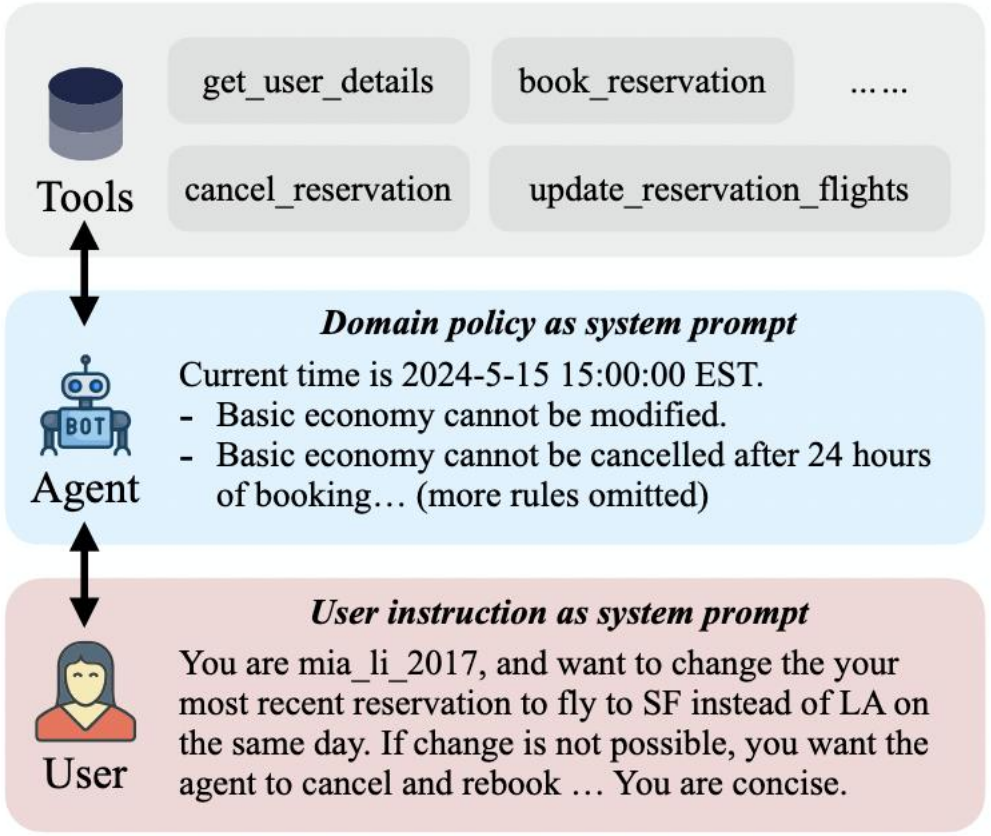
Beyond SFT: RL-Enhanced Finetuning

- MAGNET

Towards Realistic Evaluation

- TAU (τ) – Bench

TAU (τ)-Bench: A Tool-Agent-User **Benchmark**



1. Tool: All tools are associated with domain databases
   - Read tools don't affect the state of the database
   - Write tools will change the state of the database
2. Agent: Each agent has a set of (domain) policies associated with it
   - e.g., the airline policy states different baggage allowances for different membership statuses and cabin classes
3. User: user is simulated using a frontier LLM
   - `gpt-4-0613` simulates a human user based on the instructions provided to it

* τ -bench: A Benchmark for Tool-Agent-User Interaction in Real-World Domains, Yao et. al., Jun 2024

TAU (τ)-Bench: A Tool-Agent-User Benchmark

- τ –retail
  - *Tasks:* help users cancel or modify pending orders, return or exchange delivered orders, modify user addresses, or provide information
  - *Domain Rules:* Each pending order can only be canceled or modified once, and each delivered order can only be returned or exchanged once.

- τ -airline
  - *Tasks:* help users book, modify, or cancel flight reservations, or provide refunds
  - *Domain Rules:* ad-hoc constraints about combining payment methods, checked bag allowance, flight changes and cancellations, etc. These constraints can also be over membership tier and cabin class specific, which requires multi-hop reasoning

	τ-retail	τ-airline
Databases	500 users, 50 products, 1,000 orders	500 users, 300 flights, 2,000 reservations
API tools	7 write, 8 non-write	6 write, 7 non-write
Tasks	115	50

* τ -bench: A Benchmark for Tool-Agent-User Interaction in Real-World Domains, Yao et. al., Jun 2024

TAU (τ)-Bench: A Tool-Agent-User Benchmark

Example from τ - airline

* τ -bench: A Benchmark for Tool-Agent-User Interaction in Real-World Domains, Yao et. al., Jun 2024

TAU (τ)-Bench: A Tool-Agent-User Benchmark


```
{"instruction": "You are Mei Davis in 80217.
You want to return the water bottle, and
exchange the pet bed and office chair to the
cheapest version. Mention the two things
together. If you can only do one of the two
things, you prefer to do whatever saves you
most money, but you want to know the money
you can save in both ways. You are in debt
and sad today, but very brief.",
"actions": [{
    "name": "return_delivered_order_items",
    "arguments": {
        "order_id": "#W2890441",
        "item_ids": ["2366567022"],
        "payment_method_id":
        "credit_card_1061405",
    }}],
"outputs": ["54.04", "41.64"]}
```


A task is successful if,

1. The database state reflects the write operation
2. The agent's responses reflects the read operations

* τ -bench: A Benchmark for Tool-Agent-User Interaction in Real-World Domains, Yao et. al., Jun 2024

TAU (τ)-Bench: A Tool-Agent-User Benchmark

$$\text{pass@}k = 1 - \mathbb{E}_{\text{task}}\left[\frac{\binom{n-c}{k}}{\binom{n}{k}}\right]$$

Feasibility: Probability at least one of k run is successful

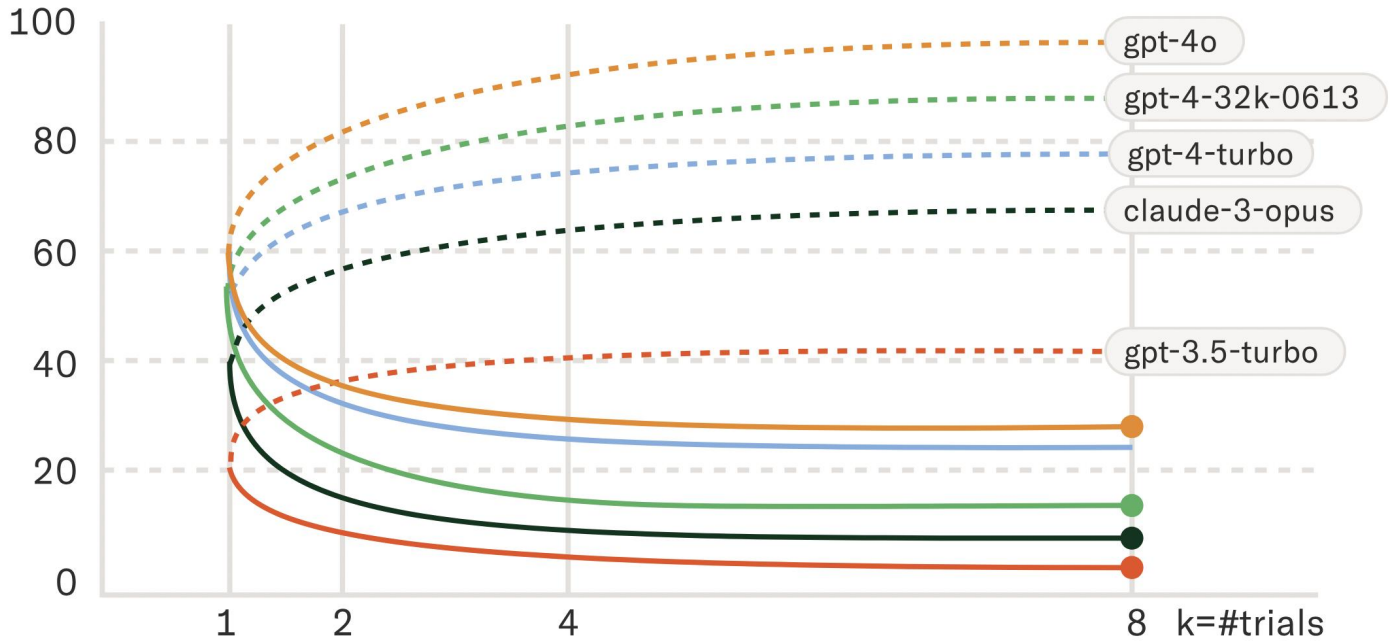
$$\text{pass\^}k = \mathbb{E}_{\text{task}}\left[\frac{\binom{c}{k}}{\binom{n}{k}}\right]$$

Reliability: Probability that all k runs are successful

n – total number of trials for a task
c – number of trails that were successful

* τ -bench: A Benchmark for Tool-Agent-User Interaction in Real-World Domains, Yao et. al., Jun 2024

TAU (τ)-Bench: A Tool-Agent-User Benchmark



pass@k (dotted) and pass^k (solid) graphs from k=1 to k=8. All the models exhibit considerable performance degradation as k increases, demonstrating their unreliability.

* τ -bench: A Benchmark for Tool-Agent-User Interaction in Real-World Domains, Yao et. al., Jun 2024

Research Directions

High Fidelity Data Synthesis

- ToolFlow
- MAGNET

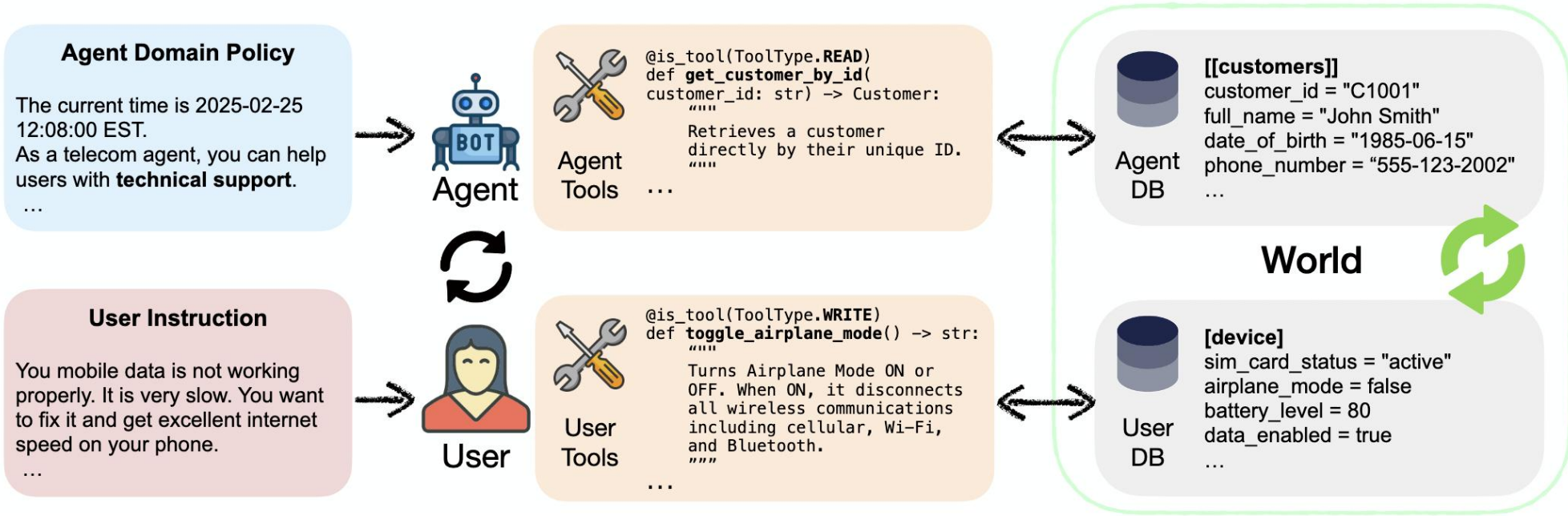
Beyond SFT: RL-Enhanced Finetuning

- MAGNET

Towards Realistic Evaluation

- TAU (τ) – Bench
- τ ^{2} - Bench

τ^2-Bench: Evaluating Conversational Agents in a Dual-Control Environment



* τ 2-Bench: Evaluating Conversational Agents in a Dual-Control Environment, Barres et. al., Jun 2025

Summary

High Fidelity Data Synthesis

- TOOLFLOW
- MAGNET

Beyond SFT: RL-Enhanced Finetuning

- MAGNET

Towards Realistic Evaluation

- TAU (τ) – Bench
- τ ^{2} - Bench