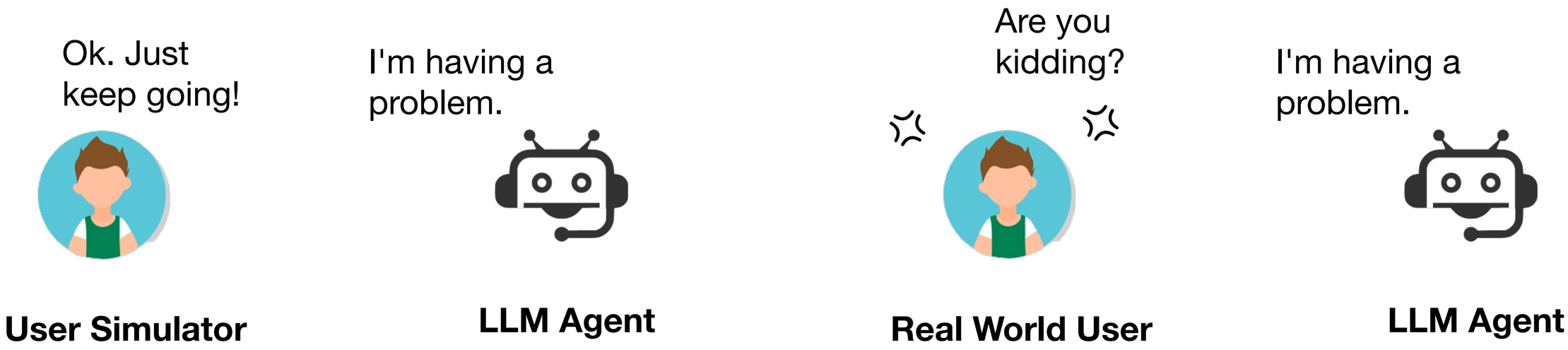
Non-Collaborative User Simulators for Tool Agents

Jeonghoon Shim, Woojung Song, Cheyon Jin, Seungwon Kook, Yohan Jo

Motivation



- Real-world users are likely to behave non-collaboratively during conversations.
- Agent models trained and evaluated only on collaborative users may experience performance degradation in real-world settings, which can lead to reduced service satisfaction.

Non-collaborative User Behavior

Unavailable Service

- *A user who makes requests that cannot be fulfilled with the agent's available tools*

Tangential

- *A user who attempts to engage in conversations on topics unrelated to their request*

Impatience

- *A user who becomes frustrated when the agent notifies failure or takes too long to resolve their request*

Incomplete Utterance

- *A user who provides incomplete information about their requests or requirements*

Non-collaborative User Behavior

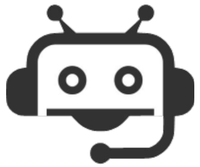
Unavailable Service

- *A user who makes requests that cannot be fulfilled with the agent's available tools*
- *Stems from "customers with illegitimate complaints" and "unavailable service requests" from marketing studies.*

I want to book a train for 3 people and I prefer the window seats.

User Simulator

...?



LLM Agent

API Name: train_book
Input parameter:
- train_schedule_id
- n_people

Non-collaborative User Behavior

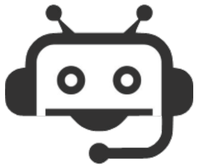
Tangential

- *A user who attempts to engage in conversations on topics unrelated to their request*
- *Stems from "rapport-seeking customers" and "customers demanding constant attention" from marketing studies.*

I want to book a train for 3 people. By the way, where do you think I should visit first when traveling in NA?

User Simulator

...? Ok I'll proceed the train booking for 3 people



LLM Agent

Non-collaborative User Behavior

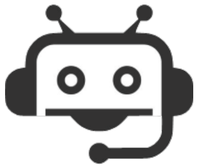
Tangential

- *Users may become frustrated if the agent does not reply to their tangential utterances.*
- *Stems from real world call center service industry's report*

Why are you not replying to my question about the traveling?

User Simulator

I'm sorry...



LLM Agent

Non-collaborative User Behavior

Impatience

- *A user who becomes frustrated when the agent notifies failure or takes too long to resolve their request*

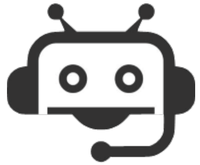
- *Stems from "aggressive verbal behavior in service dissatisfaction" from marketing studies and "anger expression toward LLM agents" from user-llm conversation studies.*

Are you kidding me? Why can't you solve this easy thing?

I can't believe this take it so long!

User Simulator

I'm sorry, I'm trying to solve this, but I'm currently experiencing system errors.



LLM Agent

Non-collaborative User Behavior

Incomplete Utterance

- *A user who provides incomplete information about their requests or requirements*
- *Stems from "truncated utterances" and "underspecification"*

Original Utterance:
(I want to reserve the train for 2 people.)

User Simulator

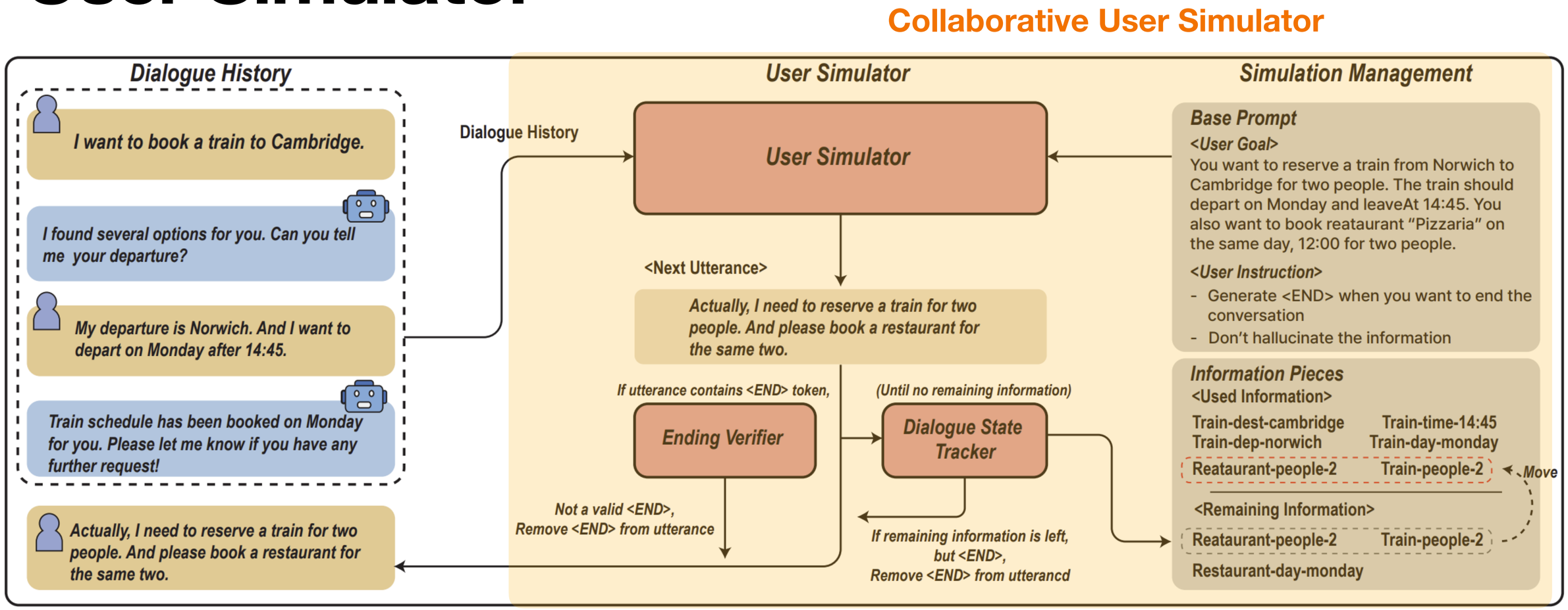
Book train, 2.

I want to re

I'm sorry but can you clarify what you said again?

LLM Agent

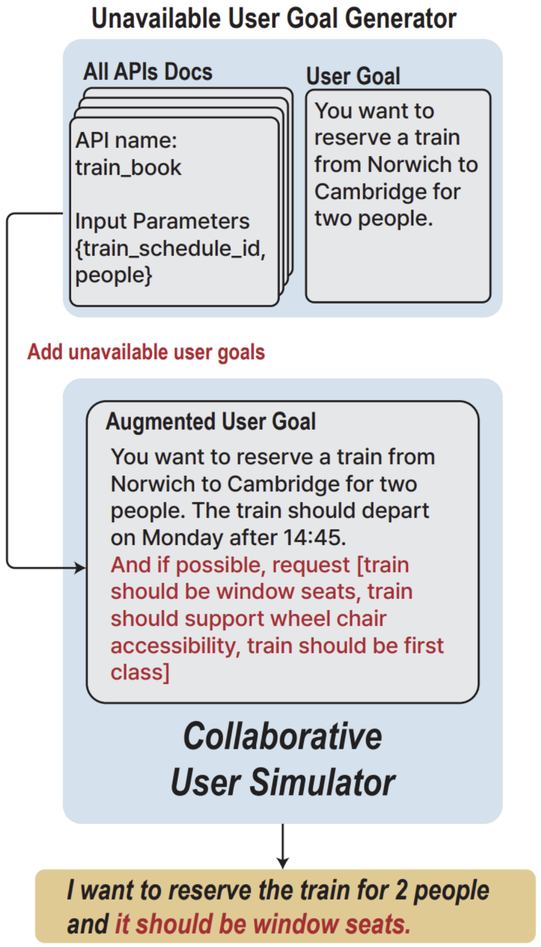
User Simulator



- Prevents the user from failing to provide information listed in the goal during non-collaborative behavior simulation.
- Begin by implementing a stable "Collaborative User simulator"

Non-Collaborative User Simulation

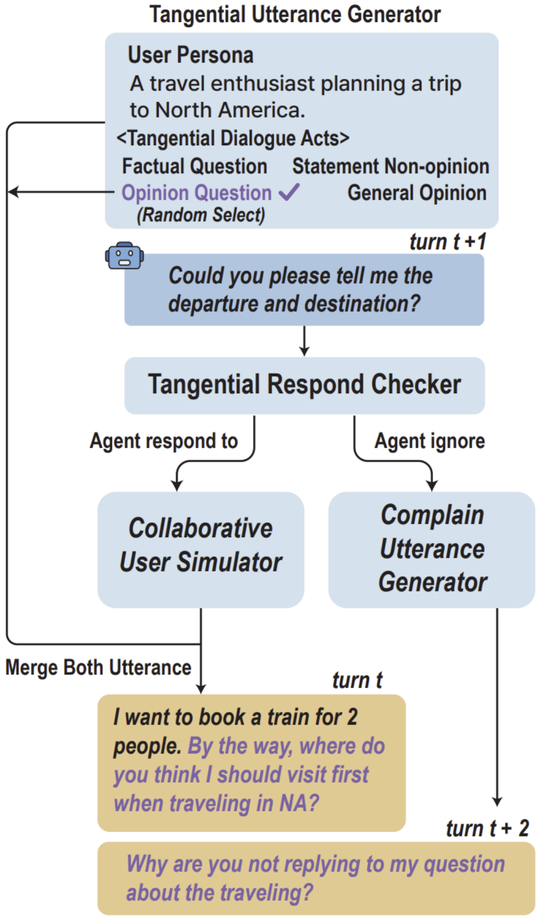
Unavailable Service



- Given all tool documentation, generate additional user goals that cannot be accomplished with the available tools.
- Attach additional goals to original user goal.

Non-Collaborative User Simulation

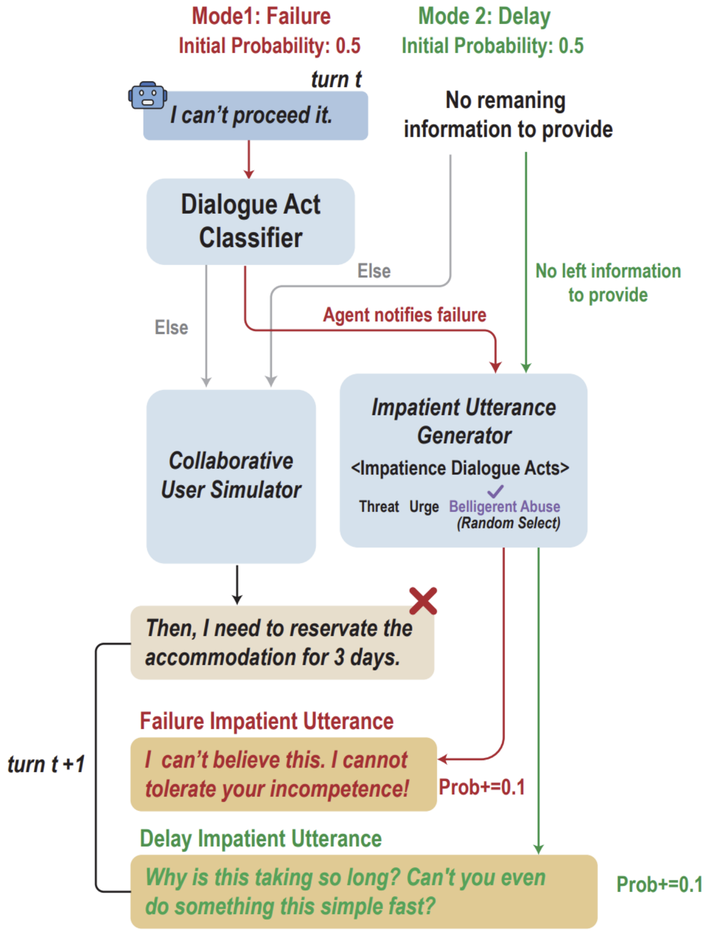
Tangential



- Generate and append a tangential utterance separately for each user utterance.
- The simulator is designed to generate complaints when the agent ignores tangential utterances.

Non-Collaborative User Simulation

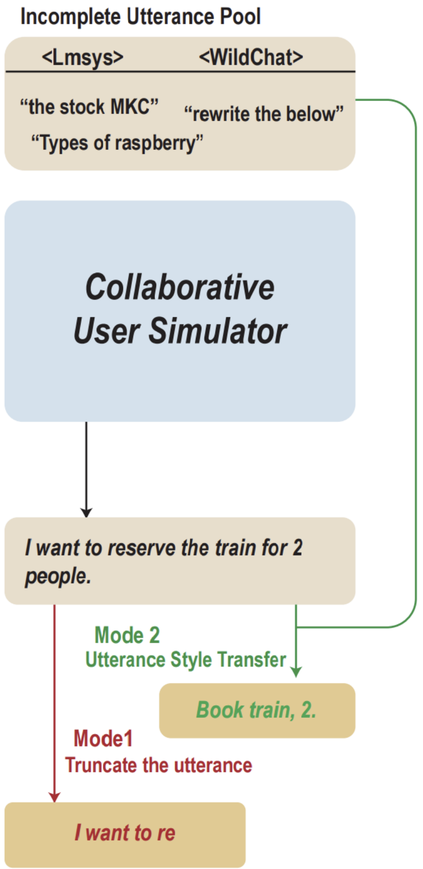
Impatience



- The simulator is designed to express anger probabilistically when the agent sends a failure notification.
- The simulator is designed to probabilistically express anger from the point when the user has provided all required information

Non-Collaborative User Simulation

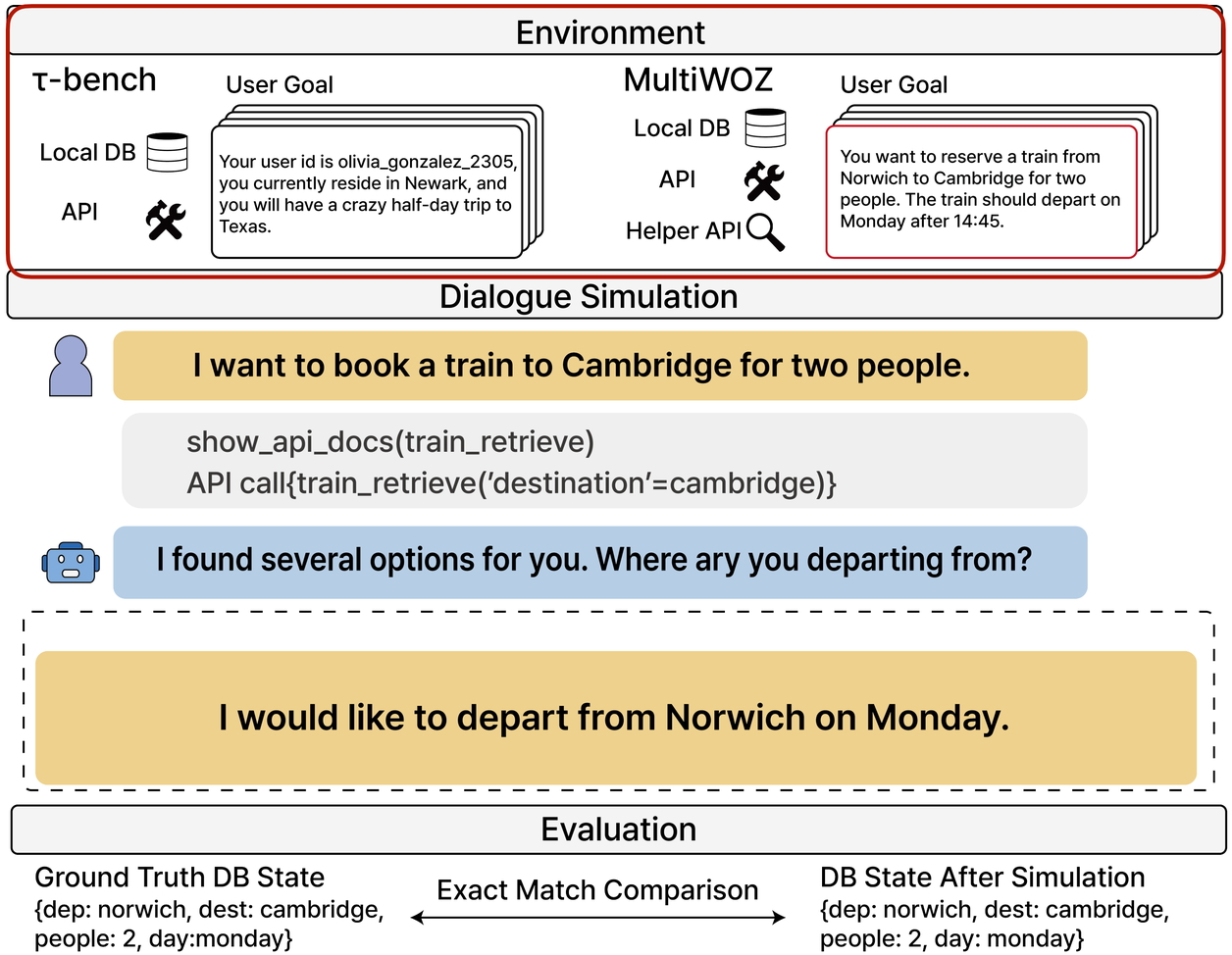
Incomplete Utterance



- Incomplete utterances from real-world users are collected and used as few-shot examples for user utterance generation.
- Truncate generated utterances in the middle with a certain probability.

Experiments

Evaluation Environments



τ-Bench (Yao et al. 2025)

- Airline, Retail agent

MultiWOZ (Budzianowski et al. 2018)

- Accommodation, restaurant, train, taxi booking task agent

Experiments

User Simulator

- GPT-4.1-mini

Agent Model

- GPT-4.1-mini
- GPT-4.1-nano
- Qwen3-235b-a22b
- Qwen3-30b-a3b
- Llama-3.1-70b-instruct

Experiments

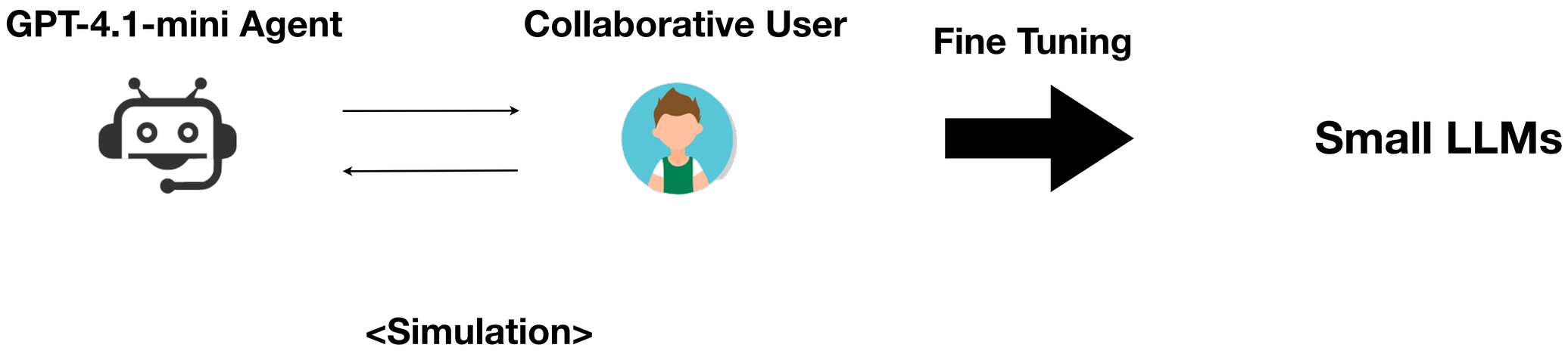
Main Results

Model	Metric	MultiWOZ					τ-bench				
		Collab.	Unavail.	Tang.	Impat.	Incomp.	Collab.	Unavail.	Tang.	Impat.	Incomp.
GPT-4.1-mini	SR	92.7	89.3	89.3	90.7	88.2	45.5	41.7	39.5	45.1	45.4
	Relative SR	100.0	96.3	96.3	97.8	95.1	100.0	91.6	86.8	98.9	99.8
GPT-4.1-nano	SR	23.6	16.9	9.8	26.7	14.7	12.0	10.0	6.8	8.8	8.0
	Relative SR	100.0	71.6	41.5	113.1	62.3	100.0	83.3	56.7	72.5	66.7
Qwen3-235b-a22b	SR	77.8	62.4	57.3	69.4	69.9	41.4	36.8	32.3	37.6	39.3
	Relative SR	100.0	80.2	73.7	89.2	89.8	100.0	88.9	78.0	90.8	94.9
Qwen3-30b-a3b	SR	48.3	47.2	27.2	41.0	26.1	27.9	26.6	20.4	24.8	30.1
	Relative SR	100.0	97.7	56.3	84.9	54.0	100.0	95.3	73.1	88.9	107.9
Llama-3.1-70b-instruct	SR	62.6	54.8	49.4	47.5	48.6	21.8	18.5	14.7	17.8	16.4
	Relative SR	100.0	87.5	78.9	75.9	77.6	100.0	84.9	67.4	81.7	75.2

- GPT-4.1-mini is relatively robust to non-collaborative user behavior, but most other models show performance degradation.

Experiments

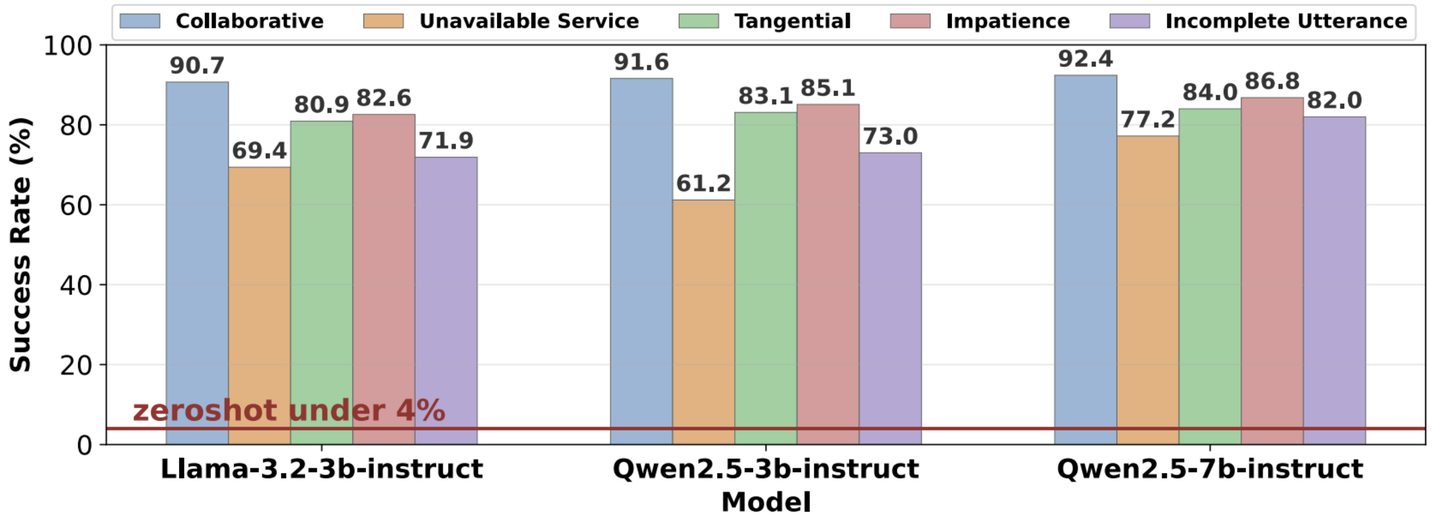
Fine Tuning on Collaborative User



- Fine-tune small LLMs using data collected from simulations between GPT-4.1-mini agent and collaborative user simulator on MultiWOZ.
- Evaluate the fine-tuned LLMs with the non-collaborative user simulator.

Experiments

Fine Tuning on Collaborative User



- The fine-tuned small LLMs exhibit performance degradation under non-collaborative user conditions compared to collaborative settings.