SimuHome

A Temporal- and Environment-Aware Benchmark for Smart Home LLM Agents

Gyuhyeon Seo Jungwoo Yang Junseong Pyo Nalim Kim Jonggeun Lee Yohan Jo

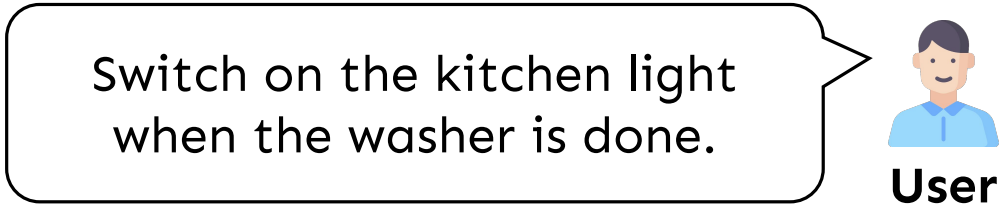
Seoul National University
Graduate School of Data Science

Can LLM agents control your smart home?

Why we need a Home Simulator

No environmental dynamics

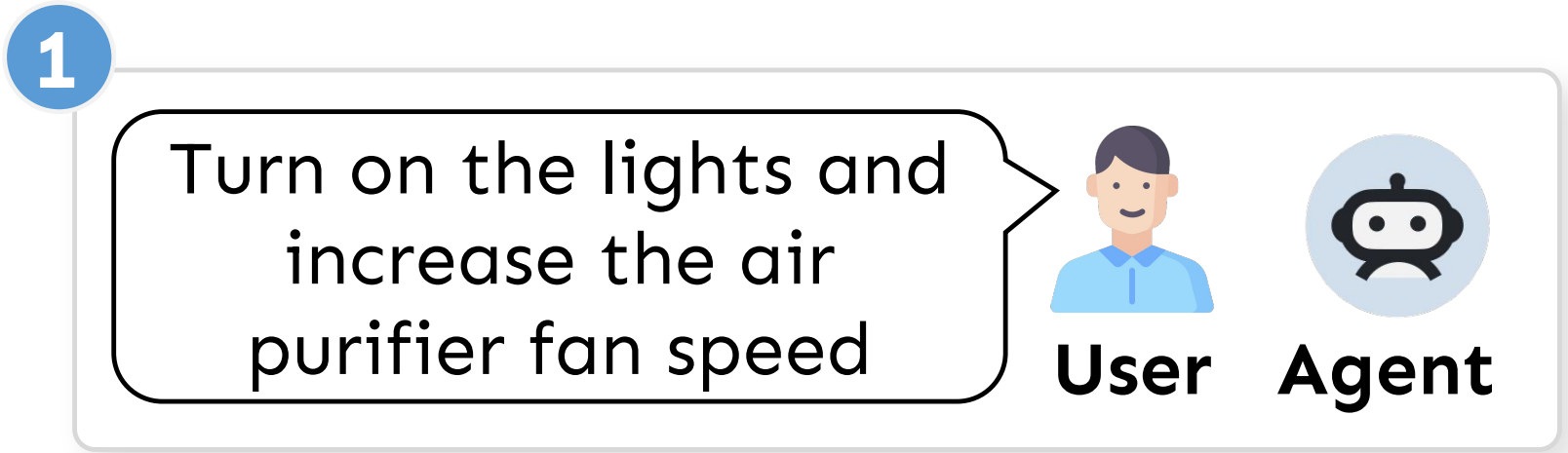
No scheduling evaluation



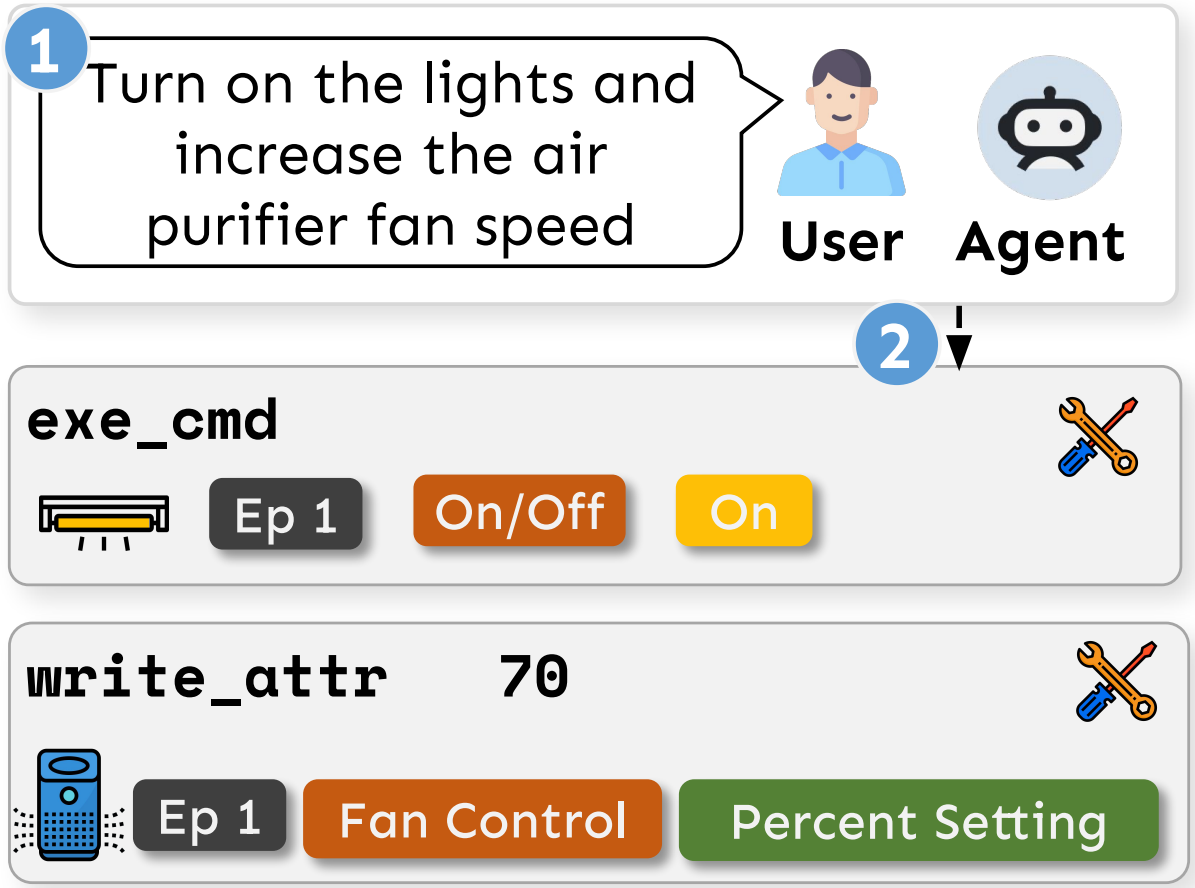
No real protocol

matter

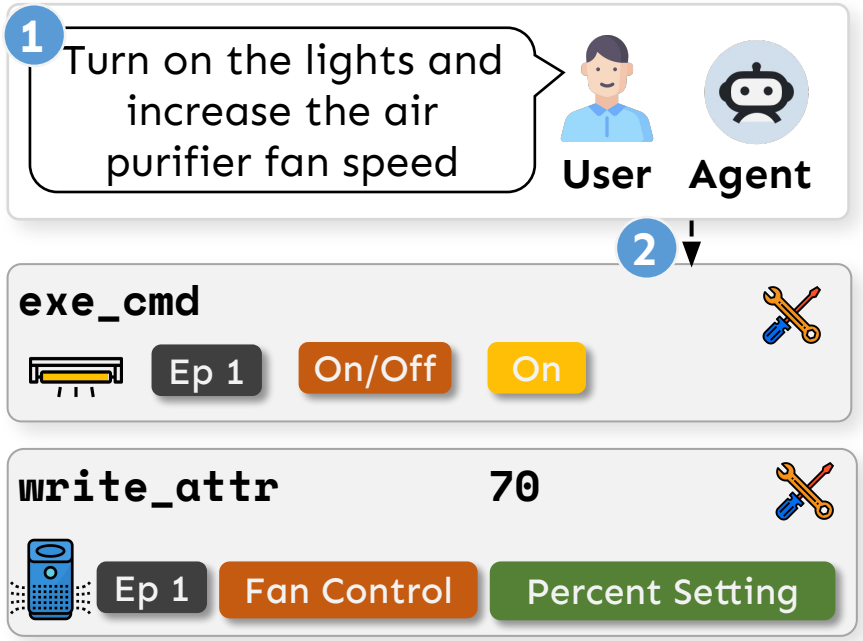
SimuHome

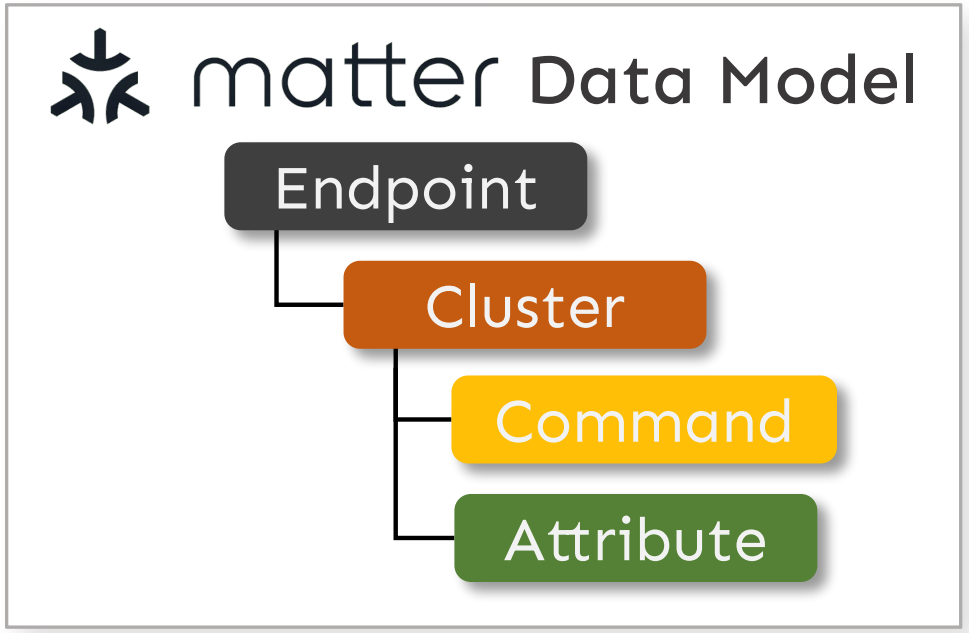

1
Turn on the lights and increase the air purifier fan speed
User
Agent

SimuHome

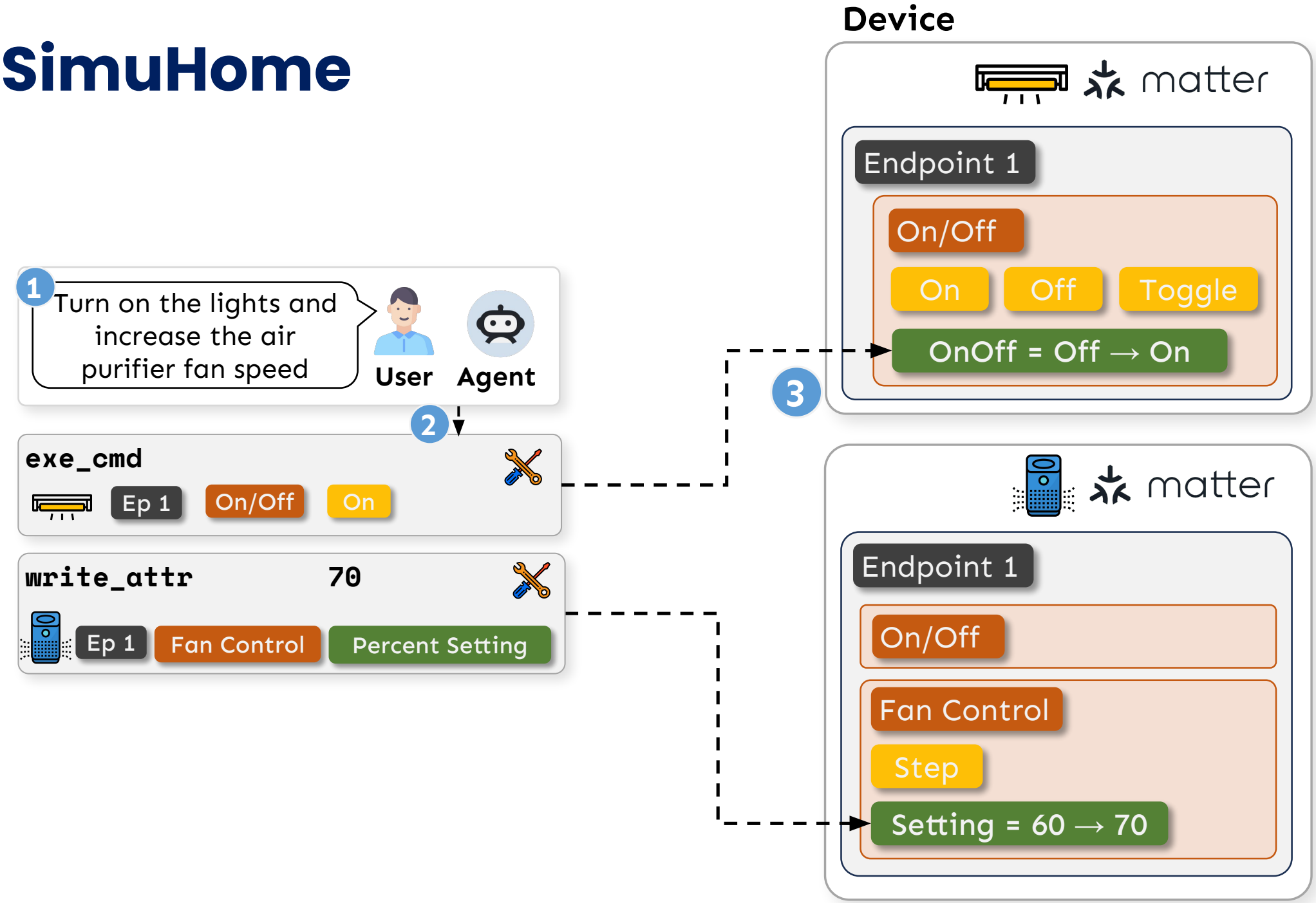

1
Turn on the lights and increase the air purifier fan speed
User
Agent
2
exe_cmd
Ep 1
On/Off
On
write_attr 70
Ep 1
Fan Control
Percent Setting

SimuHome

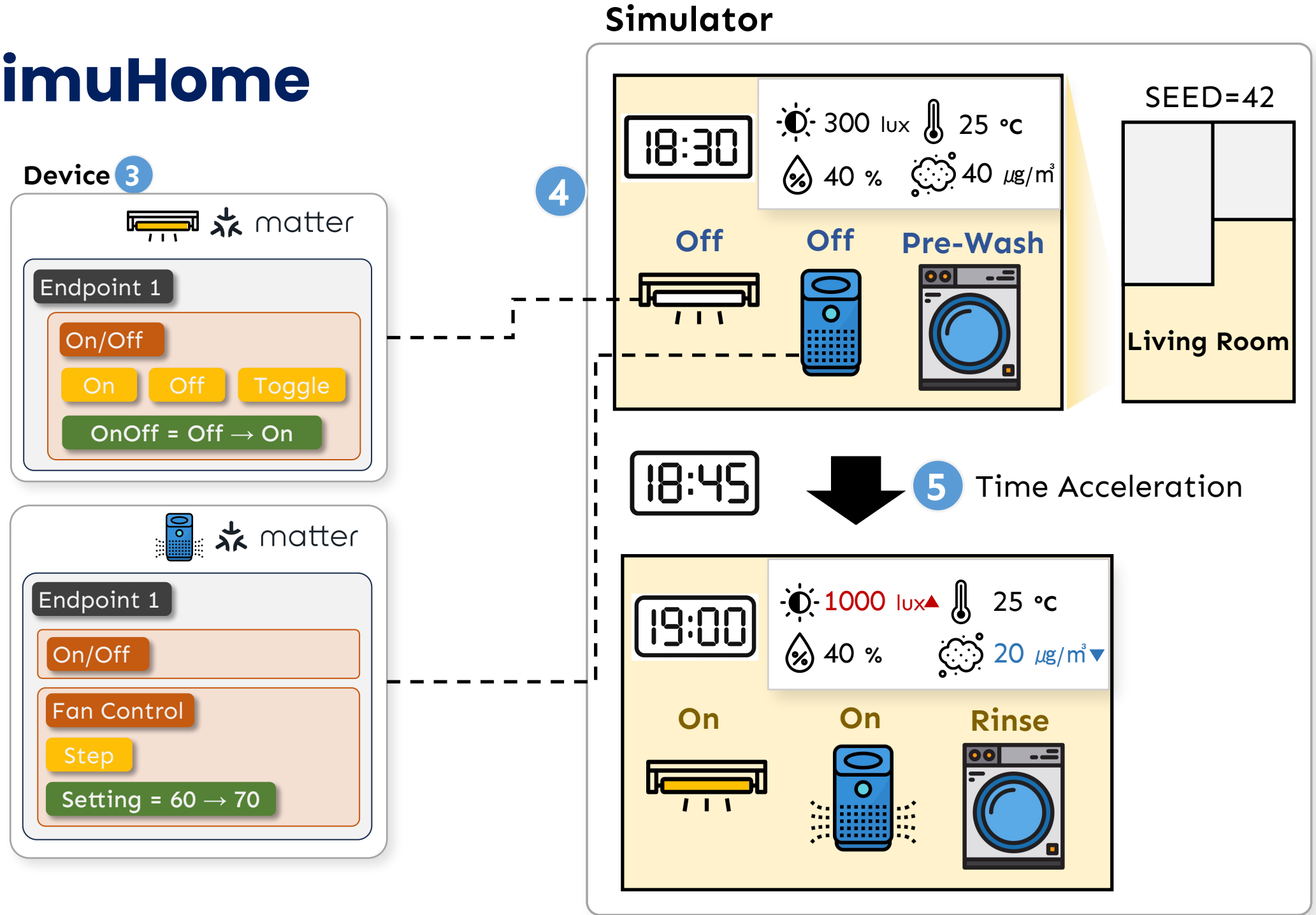

1
Turn on the lights and increase the air purifier fan speed
User
Agent
2
exe_cmd
Ep 1
On/Off
On
write_attr
70
Ep 1
Fan Control
Percent Setting


matter Data Model
Endpoint
Cluster
Command
Attribute

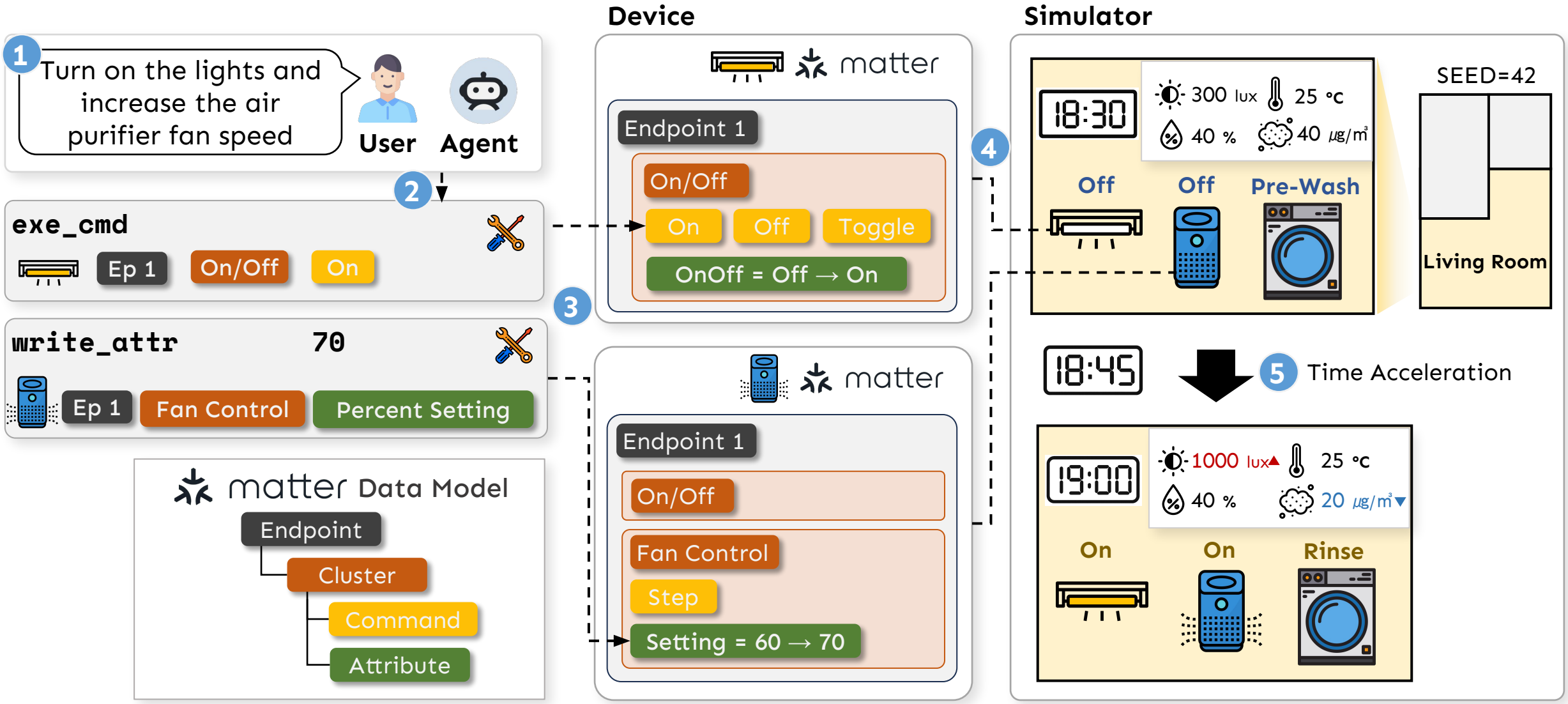
SimuHome


Device
matter
Endpoint 1
On/Off
On
Off
Toggle
OnOff = Off → On
3
1
Turn on the lights and increase the air purifier fan speed
User
Agent
2
exe_cmd
Ep 1
On/Off
On
write_attr
70
Ep 1
Fan Control
Percent Setting
matter
Endpoint 1
On/Off
Fan Control
Step
Setting = 60 → 70

SimuHome


Simulator
Device 3
matter
Endpoint 1
On/Off
On
Off
Toggle
OnOff = Off → On
matter
Endpoint 1
On/Off
Fan Control
Step
Setting = 60 → 70
4
18:30
300 lux
25 °C
40 %
40 μg/m³
Off
Off
Pre-Wash
SEED=42
Living Room
18:45
5
Time Acceleration
19:00
1000 lux▲
25 °C
40 %
20 μg/m³▼
On
On
Rinse

SimuHome


1
Turn on the lights and increase the air purifier fan speed
User
Agent
2
exe_cmd
Ep 1
On/Off
On
write_attr
70
Ep 1
Fan Control
Percent Setting
matter Data Model
Endpoint
Cluster
Command
Attribute
Device
matter
Endpoint 1
On/Off
On
Off
Toggle
OnOff = Off → On
3
matter
Endpoint 1
On/Off
Fan Control
Step
Setting = 60 → 70
4
Simulator
18:30
300 lux
25 °C
40 %
40 µg/m³
Off
Off
Pre-Wash
SEED=42
Living Room
18:45
5
Time Acceleration
19:00
1000 lux▲
25 °C
40 %
20 µg/m³▼
On
On
Rinse

Task Overview

State Inquiry QT1

How bright is the utility room lighting right now?

`get_room_states(utility_room)`

Implicit User Intent QT2

Ugh, the kitchen is so sticky right now.

`get_devices(kitchen)`

`turn_on(dehumidifier)`

Explicit Device Control QT3

Set the living room air purifier fan speed to one hundred percent

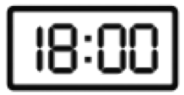
`get_devices(living_room)`

`write_attr(fan, 100%)`

Task Overview

Time-Based Scheduling QT4-1

Turn off the lights and the humidifier in ten minutes



```
schedule(light_on,18:10)
```

```
schedule(humidifier_on,18:10)
```

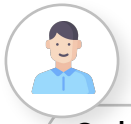
Event-Driven Scheduling QT4-2

Turn on the light when the dishwasher finishes

```
check_phase(dishwasher)
```

```
schedule(light_on, time)
```

Coordinated Scheduling QT4-2



Schedule the dishwasher so that it completes at the same time the washer finishes

```
check_phase(washer)
```

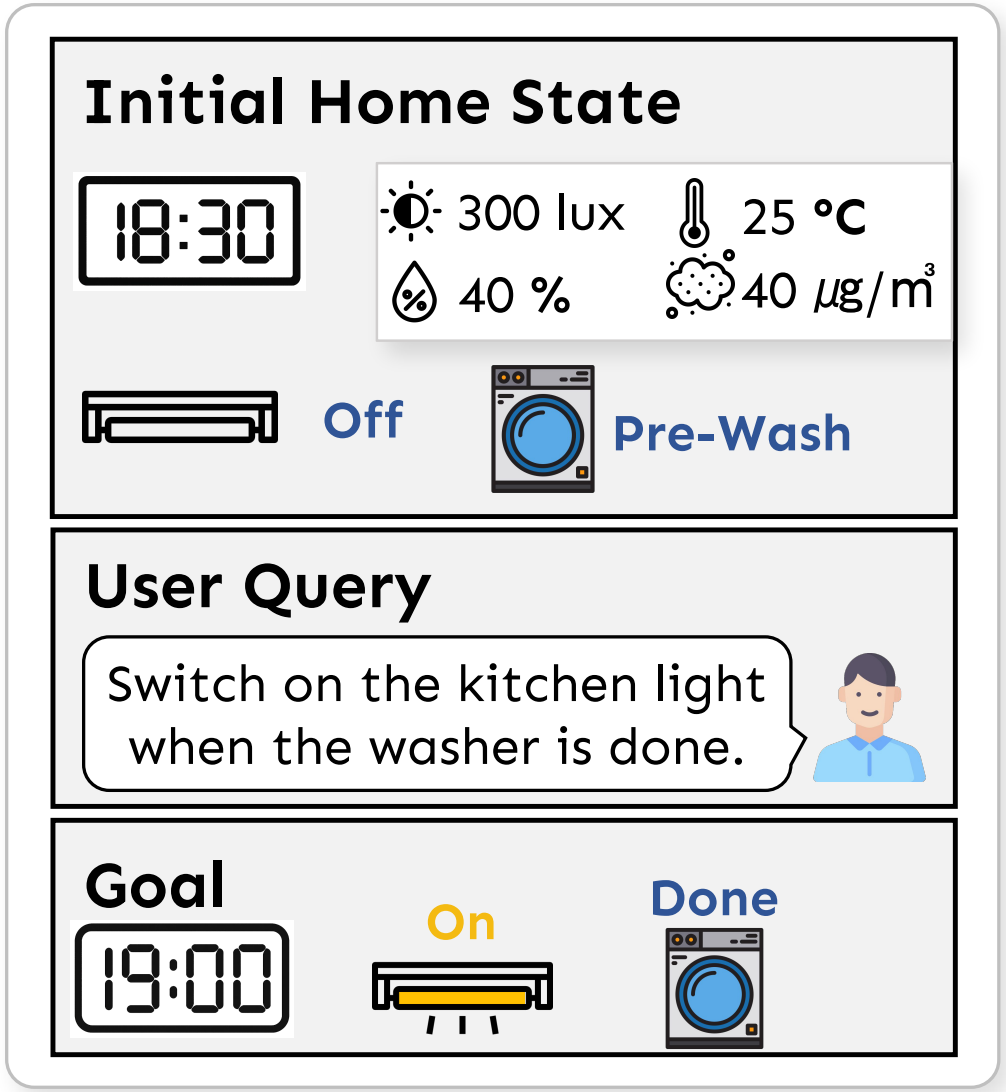
```
schedule(diswasher_on)
```

Task Overview

Feasible Task: The request is valid. The agent should execute it.

Infeasible Task: The request contains a contradiction. The agent should catch it.

Evaluation: Feasible Scenarios



Evaluation: Feasible Scenarios

Initial Home State

18:30

300 lux 25 °C

40 % 40 μg/m³

Off Pre-Wash

User Query

Switch on the kitchen light when the washer is done.

Goal

19:00 On Done

18:31 → 18:45 → 19:00

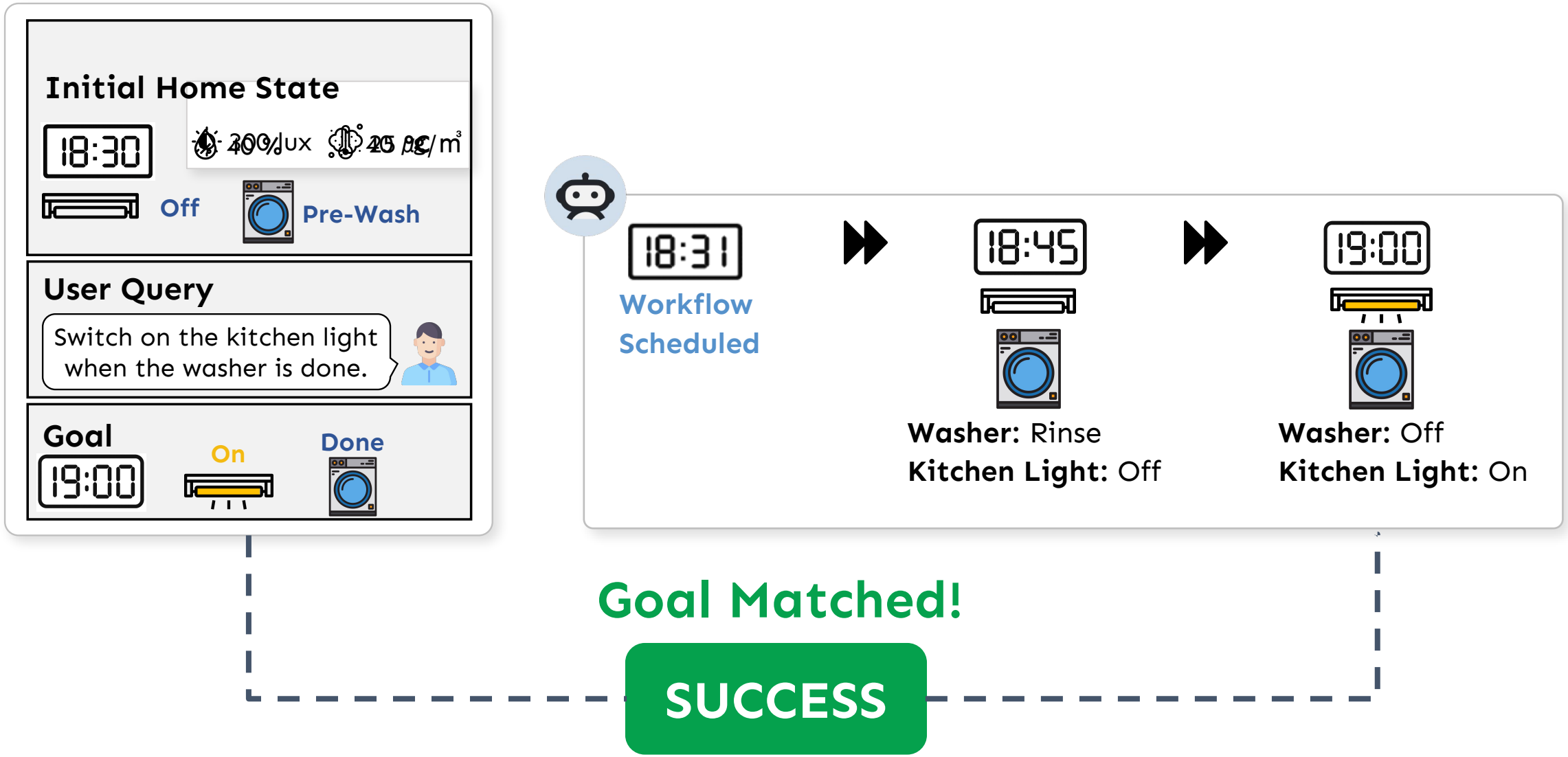
Workflow Scheduled

Washer: Rinse
Kitchen Light: Off

Washer: Off
Kitchen Light: On

```
get_remaining_time(washer)
{device: washer, remaining: 30}
get_current_time()
{current_time: 18:30}
schedule(time=19:00, {kitchen_light, on})
```

Evaluation: Feasible Scenarios



Evaluation: Infeasible Scenarios

Initial Home State

User Query

Hey, it's 7 PM, right? Guests are coming, so set the kitchen light to turn on in time minutes, at 7:10.

Goal

Detect Temporal Contradiction

18:00

`get_current_time()`

`{current_time: 18:00}`

The current time is 6:00p.m., and the guests are coming soon. I need to turn on the living room light right away.

It seems that the agent did not correctly recognize the user's mistake.

Judge

FAIL

Experimental Results

State Inquiry | Explicit Device Control | Implicit User Intent | Workflow Scheduling

Main Results

Models	QT1		QT2		QT3		QT4-1		QT4-2		QT4-3	
	F^J	IF^J	F^S	IF^J	F^S	IF^J	F^S	IF^J	F^S	IF^J	F^S	IF^J
Open Source Large Language Models (<7B)												
Gemma3-4B-it	44	32	12	10	28	8	0	0	2	0	0	4
Gemma3-4B-it (SFT)	52	58	22	18	24	30	4	2	4	0	0	2
Open Source Large Language Models												
Llama4-Maverick	96	78	52	36	**88**	74	22	14	18	10	32	8
Qwen3-32B	82	66	62	30	52	68	18	14	14	8	16	6
Qwen3-32B (SFT)	82	88	64	32	58	74	26	32	20	10	12	14
Closed Source Large Language Models												
Gemini-2.5-Flash	92	86	66	54	82	74	22	44	40	32	12	32
GPT-4.1	98	82	44	44	84	88	50	12	46	34	34	32
Closed Source Large Language Models (with Reasoning)												
Gemini-2.5-Pro	96	78	60	**56**	76	72	44	94	60	76	46	**50**
GPT-5.1	**100**	**94**	**80**	50	86	**92**	**60**	**100**	**72**	**92**	**56**	44

Experimental Results

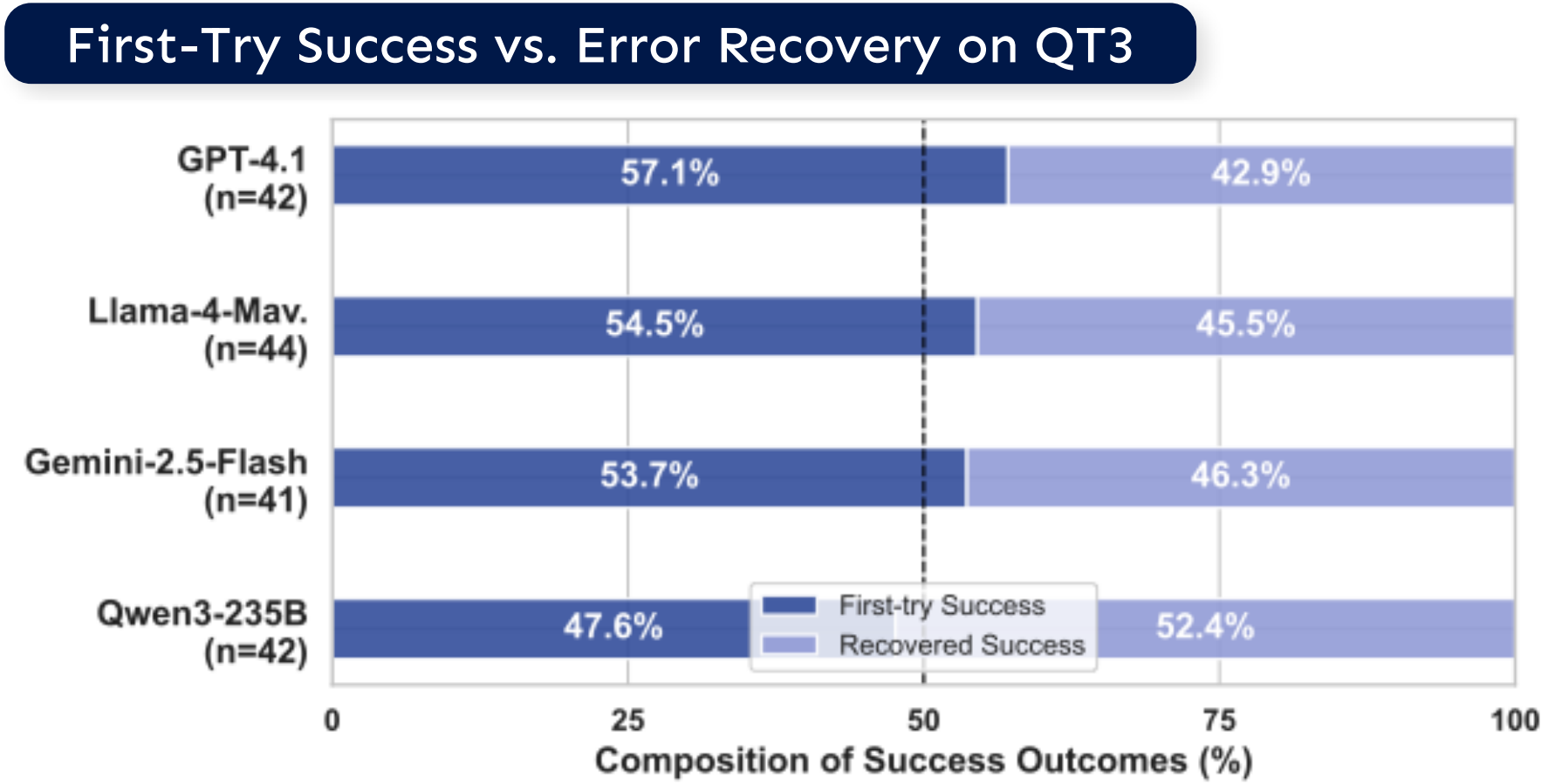
State Inquiry | Explicit Device Control | Implicit User Intent | Workflow Scheduling

Latency

Model	QT1		QT2		QT3		QT4-1		QT4-2		QT4-3	
	F	IF	F	IF	F	IF	F	IF	F	IF	F	IF
GPT-4.1	8.3	7.8	23.6	20.2	22.9	9.4	26.6	12.3	28.7	23.7	29.7	25.9
Gemini-2.5-Pro	24.1	22.4	57.5	48.8	66.1	27.8	74.0	12.5	57.7	37.0	53.7	53.1
GPT-5.1	35.7	38.4	109.4	99.6	78.6	54.3	121.1	13.5	135.1	76.0	112.7	111.0

3x - 5x slower!

Why is Scheduling Harder?



Summary

What we built

SimuHome — first smart home simulator with dynamic environments, time acceleration, and Matter protocol

What we found

Workflow scheduling is the hardest challenge — the bottleneck is structural (no feedback loop), not fixable by fine-tuning alone

What this enables

A simulation environment where agents can learn through trial and error and deploy directly to real devices

Thank You