Data Efficient Paradigms for Personalized Assessment of Taskable AI Systems

Pulkit Verma | Thesis Advisor: Siddharth Srivastava | Arizona State University | verma.pulkit@asu.edu

How would a non-expert assess the limits and capabilities of an AI system?

Introduction

Objective: Learn an interpretable model of a black-box agent by interrogating it.

Key technical challenge:

- Which sequence of queries to ask?

Abstraction in Space of Models

Abstracted model

```
(:action load_truck
 :parameters (?package ?truck ?location)
 :precondition (and (at ?truck ?location)
                (+/-/∅) (at ?package ?location))
 :effect (and (not (at ?package ?location))
              (in ?package ?truck)))
```

This predicate can appear in three forms:

- positive
- negative
- absent

abstraction ↑

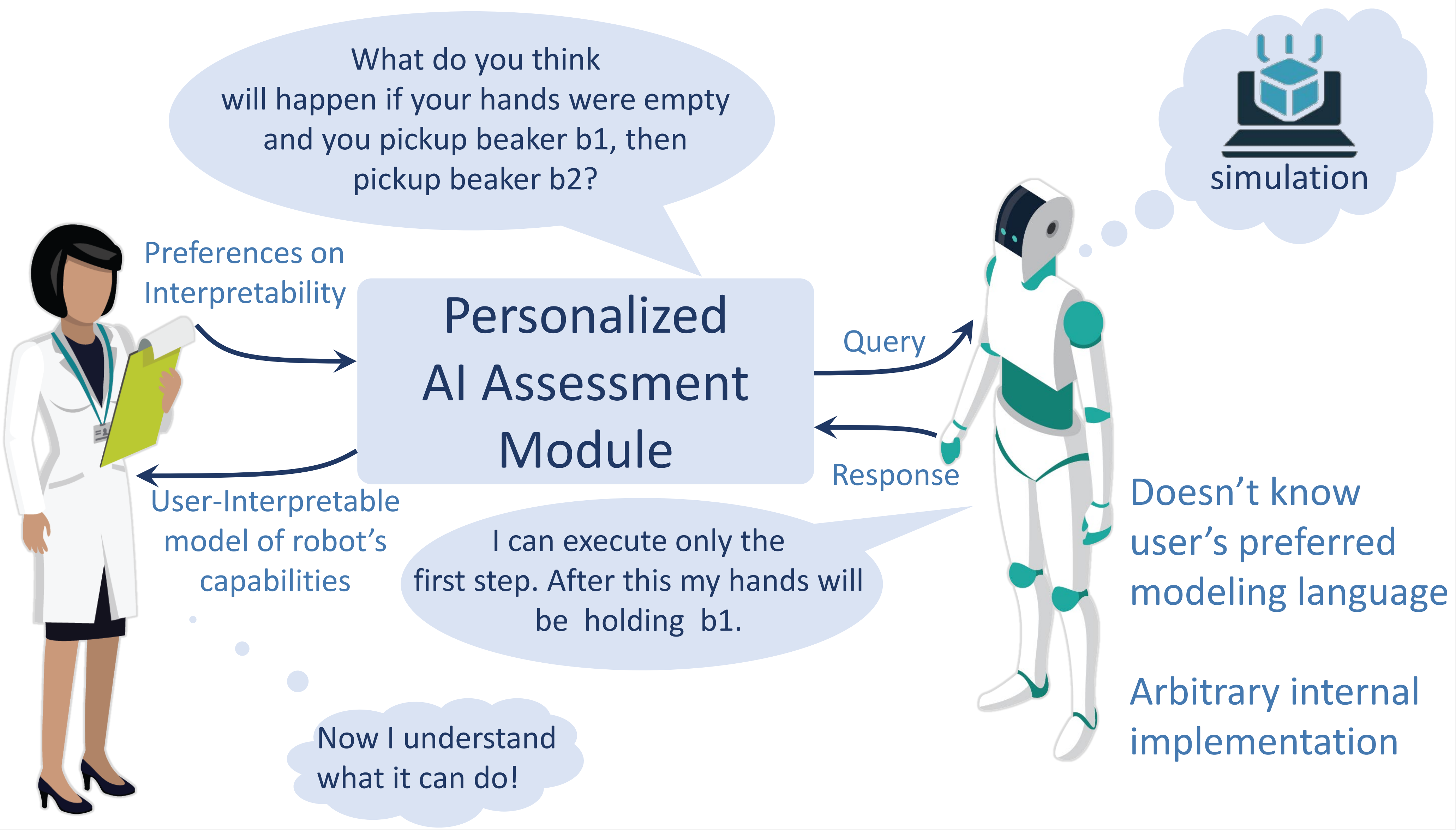
Concrete model

```
(:action load_truck
 :parameters (?package ?truck ?location)
 :precondition (and (at ?truck ?location)
                    (at ?package ?location))
 :effect (and (not (at ?package ?location))
              (in ?package ?truck)))
```

Salient Features

- Efficiently learns **causally correct** model of an AI agent's capabilities in STRIPS-like form.
- Needs no prior knowledge of the agent model.
- Only requires an agent to have rudimentary query answering capabilities.
- Queries can be answered using a simulator.

Example of Agent Interrogation



What do you think will happen if your hands were empty and you pickup beaker b1, then pickup beaker b2?

simulation

Preferences on Interpretability

Personalized AI Assessment Module

Query

Response

User-Interpretable model of robot's capabilities

I can execute only the first step. After this my hands will be holding b1.

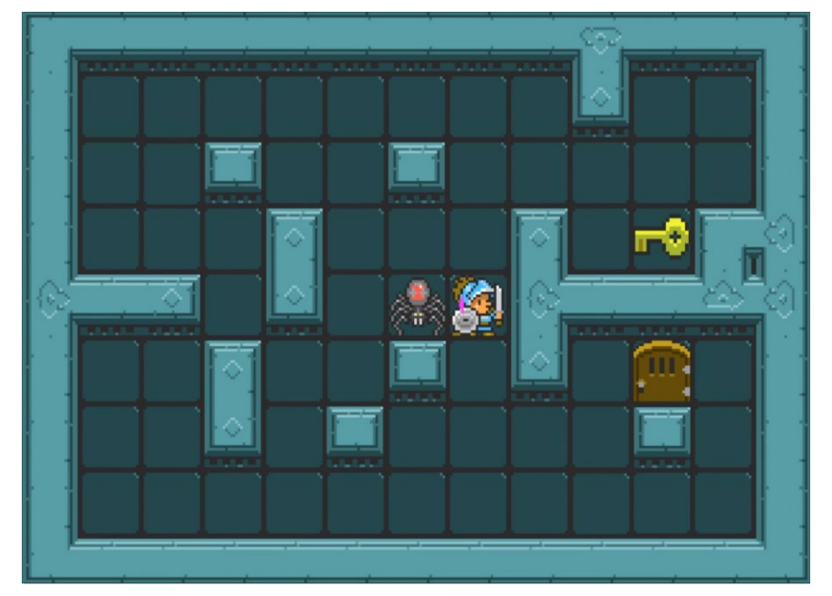
Now I understand what it can do!

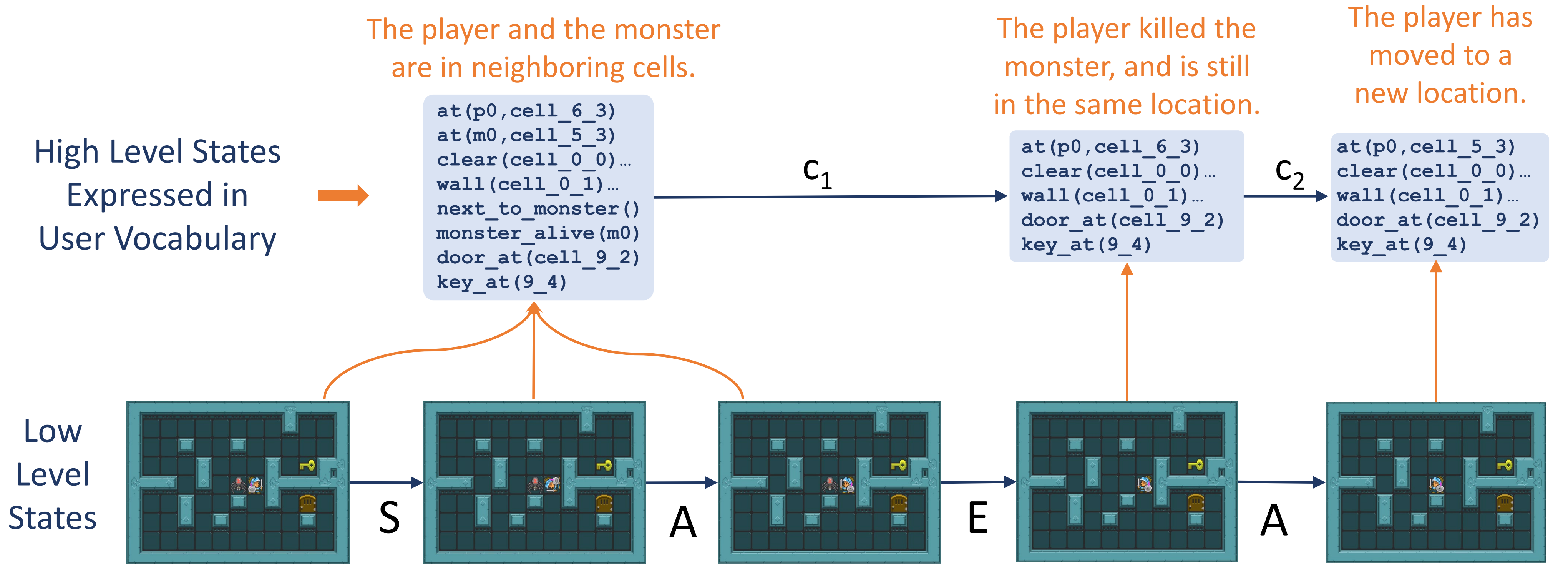
Doesn't know user's preferred modeling language

Arbitrary internal implementation

Discovering High-Level Agent Capabilities

Simulator Actions	Simulator State Representation
W	pixel_1_1(#42A8B3)
A	pixel_1_2(#42A8B3)
S	.
D	.
E	pixel_n_m(#203A3D)
Might be insufficient to understand the agent's capabilities	**Might be more expressive than what the user understands**





The player and the monster are in neighboring cells.

The player killed the monster, and is still in the same location.

The player has moved to a new location.

High Level States Expressed in User Vocabulary

```
at(p0,cell_6_3)
at(m0,cell_5_3)
clear(cell_0_0)…
wall(cell_0_1)…
next_to_monster()
monster_alive(m0)
door_at(cell_9_2)
key_at(9_4)
```

c_1

```
at(p0,cell_6_3)
clear(cell_0_0)…
wall(cell_0_1)…
door_at(cell_9_2)
key_at(9_4)
```

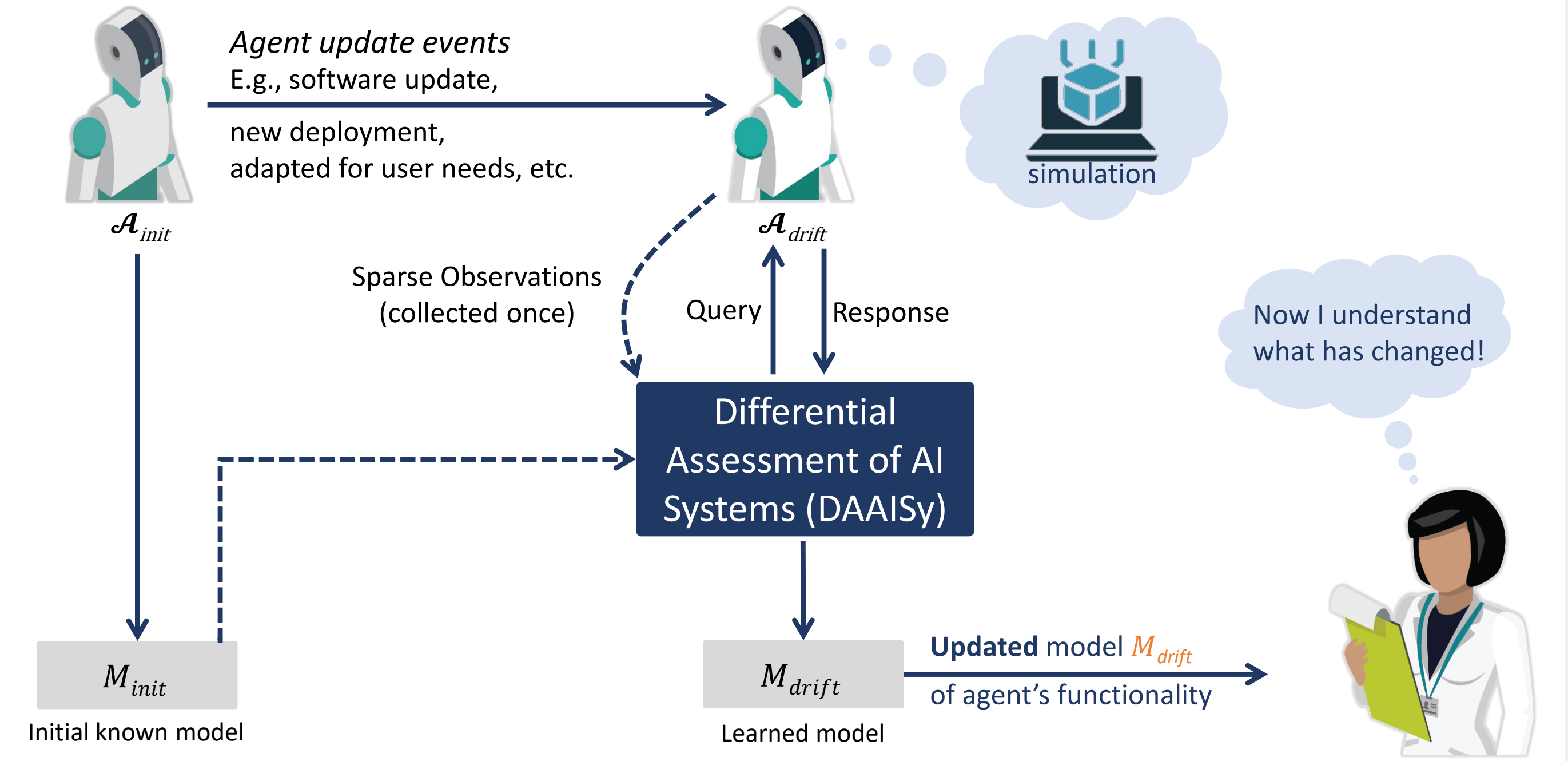
c_2

```
at(p0,cell_5_3)
clear(cell_0_0)…
wall(cell_0_1)…
door_at(cell_9_2)
key_at(9_4)
```

Low Level States

Differential Assessment

- Assess and learn model of true functionality of an adaptive black-box AI agent that has drifted from its previously known functionality.
- Identify what changed and how it changed?



Agent update events
E.g., software update,
new deployment,
adapted for user needs, etc.

simulation

Sparse Observations (collected once)

Query

Response

Differential Assessment of AI Systems (DAAISy)

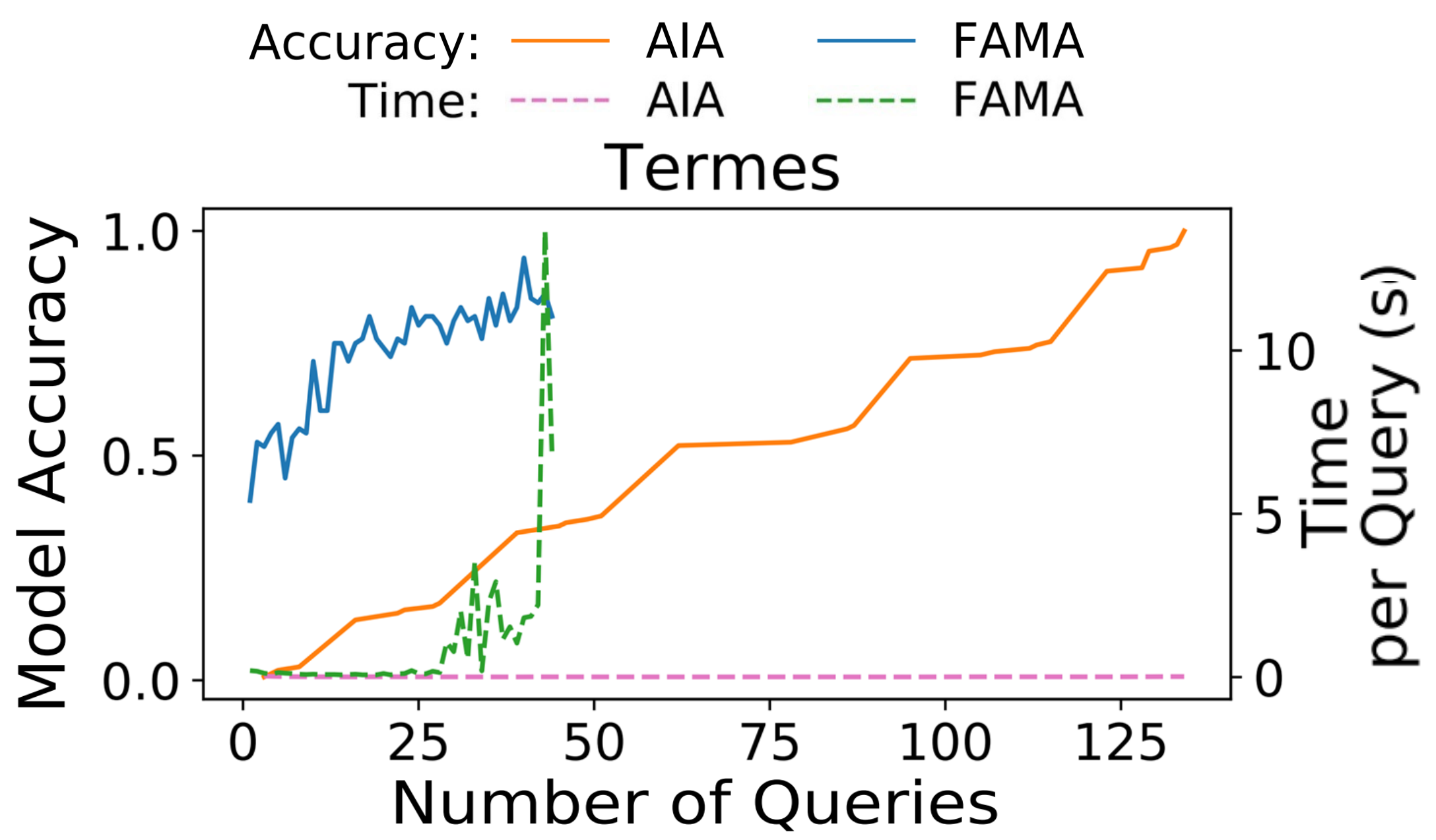
Now I understand what has changed!

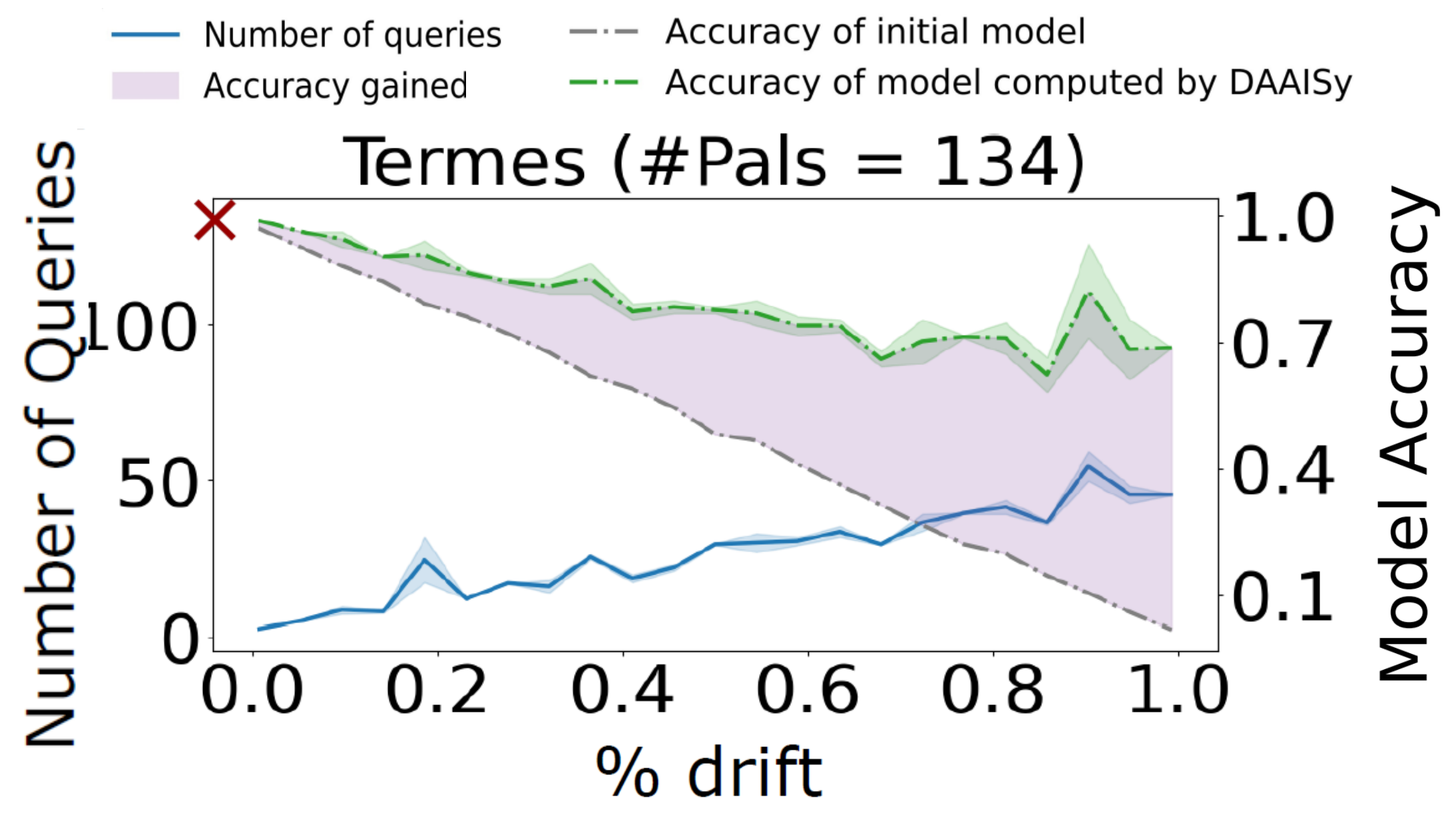
M_{init} Initial known model

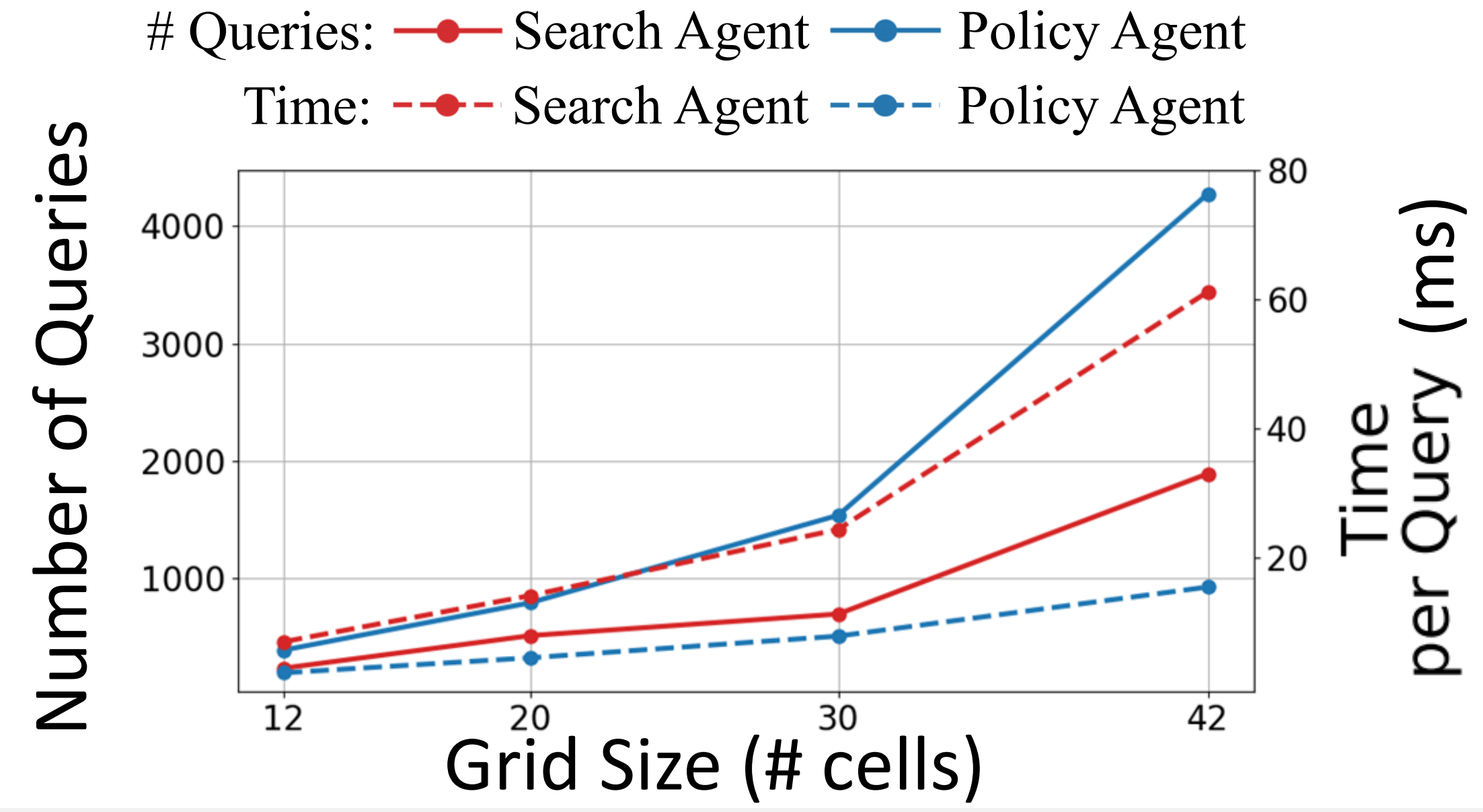
M_{drift} Learned model

Updated model M_{drift} of agent's functionality

Results







- AIA efficiently derives interpretable agent models for a range of agents.
- AIA is much faster than state of the art methods for deriving models based on passive observations.
- AIA offers better convergence guarantees.
- DAAISy can learn drifted model faster than learning from scratch using AIA.
- Policy agents take more queries to learn the agent model but learns the model faster.
- Learns all high-level actions correctly that are seen in low-level observations.

Refer to the papers for detailed results

bit.ly/3p4cVRu

bit.ly/3so0nrx

bit.ly/3theuA9