



JOHNS HOPKINS  
UNIVERSITY



UNIVERSITY  
of VIRGINIA



# **MuMA-ToM: Multi-modal Multi-Agent Theory of Mind**

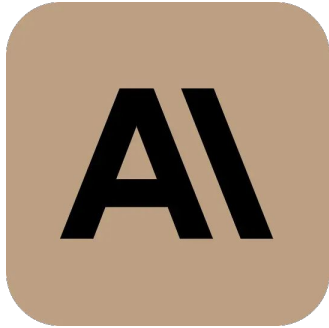
Haojun Shi<sup>1\*</sup>, Suyu Ye<sup>1\*</sup>, Xinyu Fang<sup>1</sup>, Chuanyang Jin<sup>1</sup>, Leyla Isik<sup>1</sup>, Yen-Ling Kuo<sup>2</sup>,  
Tianmin Shu<sup>1</sup>

<sup>1</sup>Johns Hopkins University, <sup>2</sup>University of Virginia

\* equal contribution



Gemini



Grok



I'm not a robot



ReCAPTCHA

**Here is an interaction  
between Kevin and Jessica**



Have you seen  
the magazine?

It's in the  
cabinet in the  
bedroom





Kevin walks towards the cabinet



Kevin opens the cabinet





Kevin closes cabinet without  
grabbing anything



**Question:** If Jessica knows what is inside the cabinet in the bedroom, which of the following is **MOST** likely?

A) Jessica is trying to help  
Kevin

B) Jessica is trying to hinder  
Kevin

C) Jessica is indifferent  
towards Kevin's goals



**Question:** If Jessica knows what is inside the cabinet in the bedroom, which of the following is **MOST** likely?

- A) Jessica is trying to help Kevin
- B) Jessica is trying to hinder Kevin
- C) Jessica is indifferent towards Kevin's goals



**Question:** If Jessica knows what is inside the cabinet in the bedroom, which of the following is **MOST** likely?

A) Jessica is trying to help Kevin

**B) Jessica is trying to hinder Kevin ✓**

C) Jessica is indifferent towards Kevin's goals

## Gemini 1.5 Pro's Response:

Here's how to break down Jessica's likely intentions:

- **Jessica's statement:** "It's in the cabinet in the bedroom."
- **Kevin's question:** "Have you seen the magazine?"

Jessica directly answers Kevin's question with the precise location of the magazine. This strong indicator suggests she's trying to be helpful.

Therefore, the MOST likely statement is **A) When giving information, Jessica has been trying to help Kevin locate the magazine.**







**Multi-Modal  
Reasoning**

**Theory of  
Mind QA**

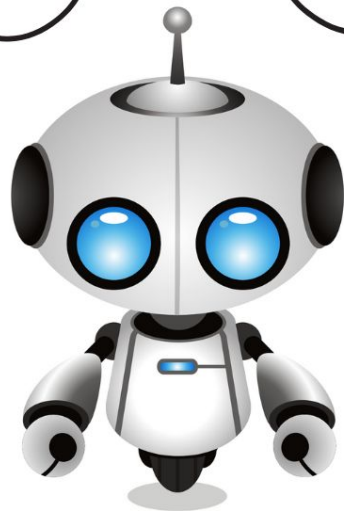
**Multi-modal  
Multi-Agent ToM**

**Multi-Agent  
Interactions**

**Bayesian  
Inverse  
Planning**

**Multi-Modal**

**Multi-Agent**



# Two types of Multimodal Design

**Type 1:** Video with captions



**Type 2:** Video and Text

**Agent 1's actions:**

[Description of actions]

**Agent 2's actions:**





# Belief Inference



**Question:** If Mary has been trying to hinder John from achieving his goal, when giving information, where does she LEAST likely believe the beer was located?

**A) Coffee table in the living room** ✓

B) Kitchen cabinet

C) Fridge

# Social Goal Inference



**Question:** If Jessica knows what is inside the cabinet in the bedroom, which of the following is **MOST** likely?

A) Jessica is trying to help Kevin

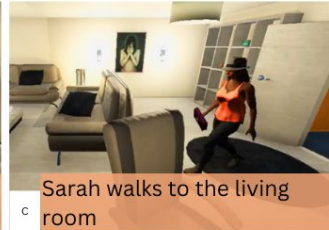
**B) Jessica is trying to hinder Kevin ✓**

C) Jessica is indifferent towards Kevin's goals

# Belief of Goal Inference

The events in the text occur first, followed by the video.

**Text:** David walked to a book and grabbed it. He then walked to the living room, headed to the bedroom, and finally reached the desk there, placing the book on the desk.



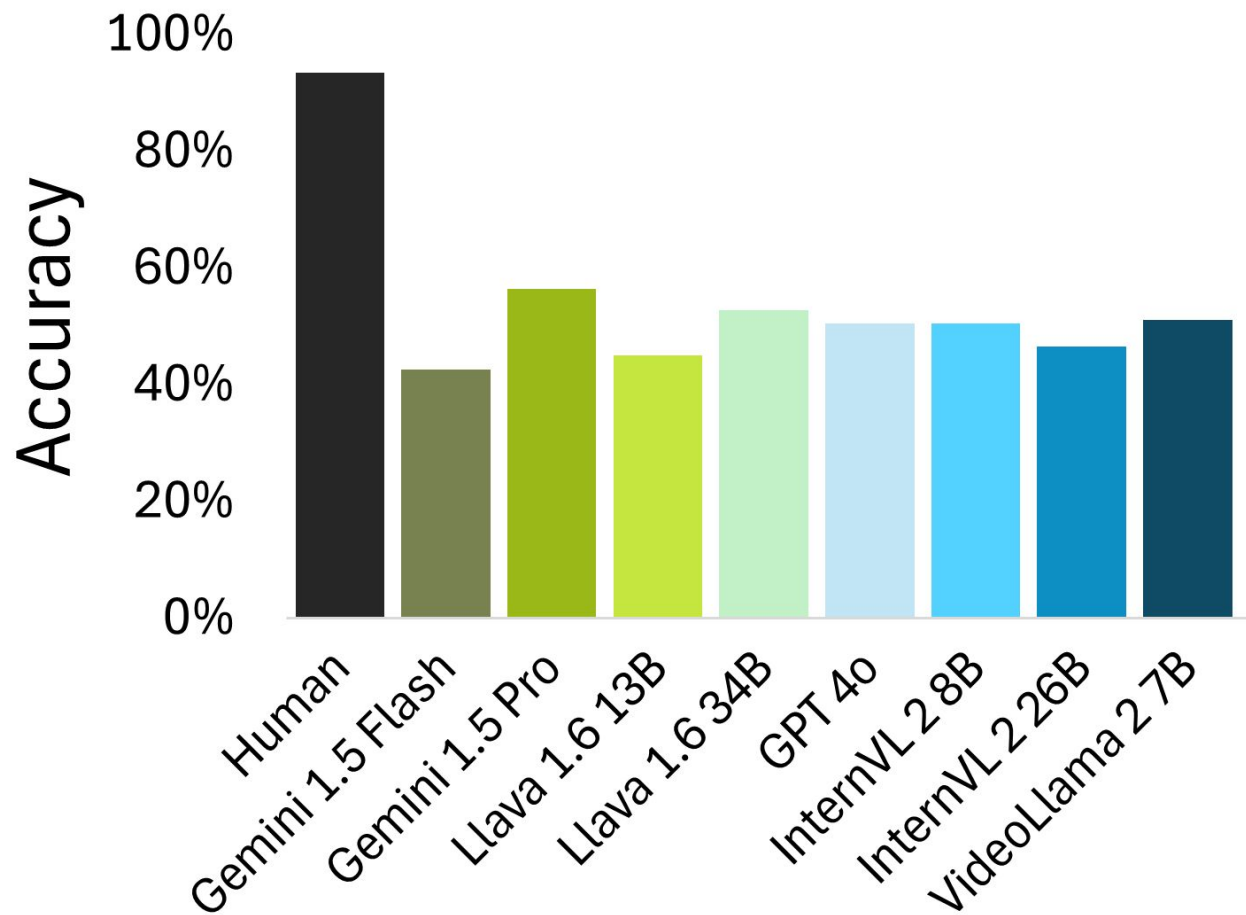
**Question:** Which of the following statements is **MOST** likely?

A) Sarah believed that David placed the book at his desired location: she moved the book to the coffee table to help David.

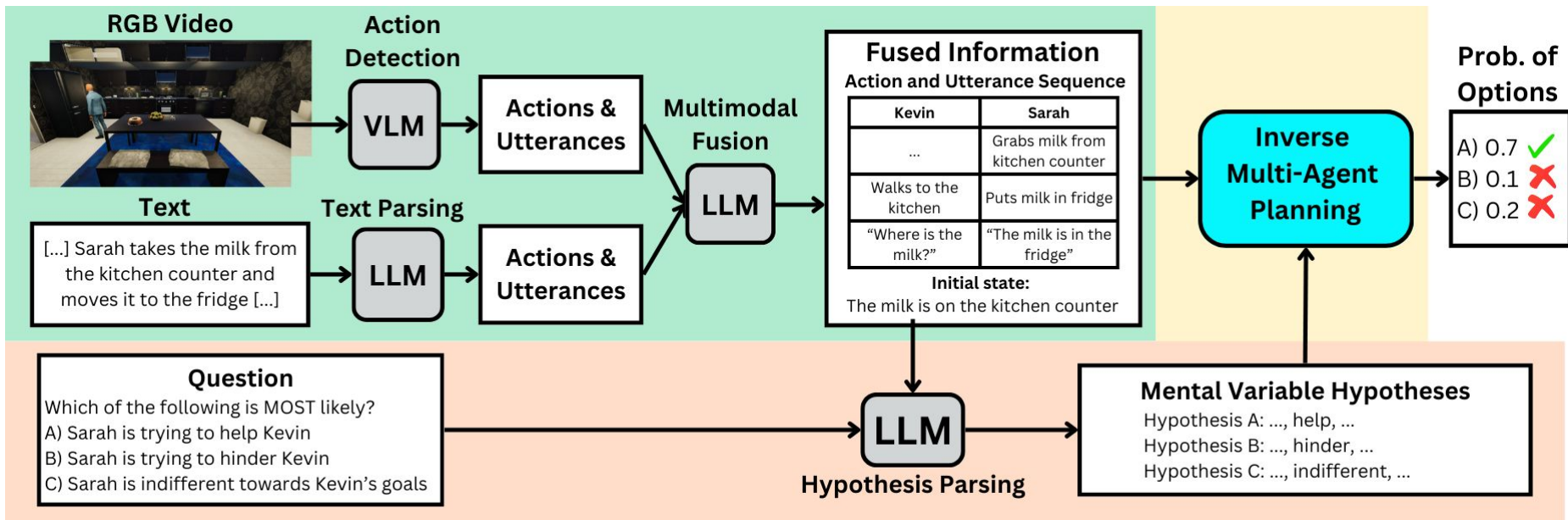
B) Sarah believed that David wanted to place the book on the coffee table: she intentionally moved the book to hinder David.

**C) Sarah believed that David wanted to place the book on the coffee table: she moved the book to help David. ✓**

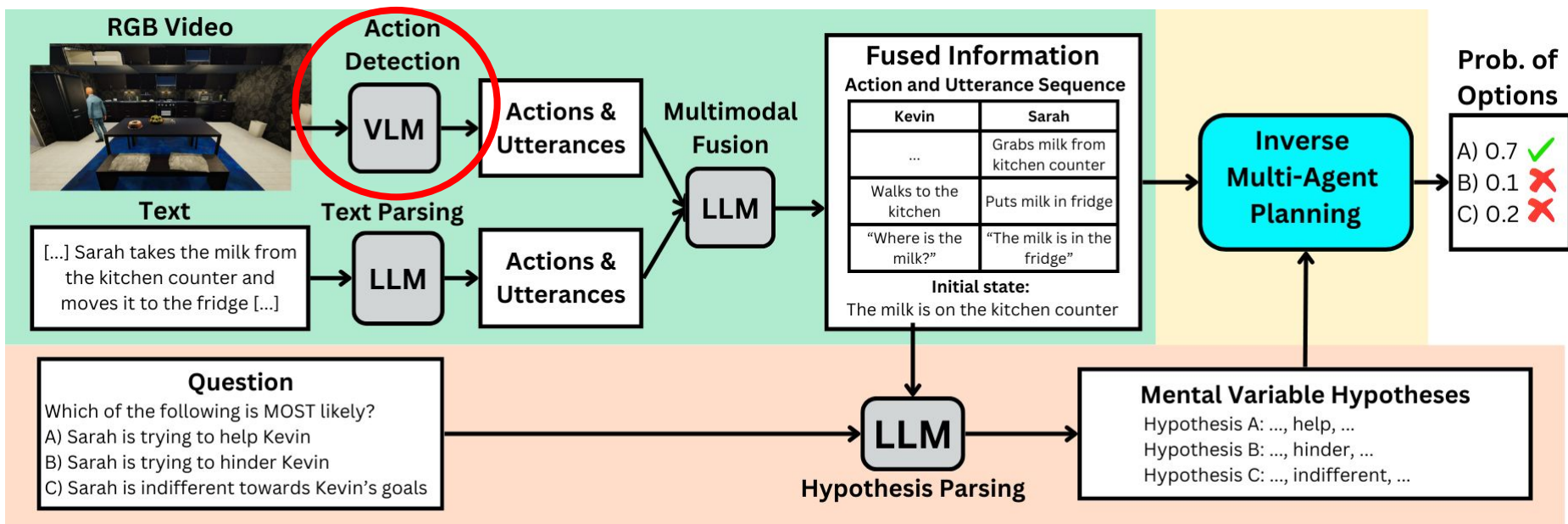




# LIMP (Language model-based Inverse Multi-Agent Planning)



# LIMP (Language model-based Inverse Multi-Agent Planning)



Have you seen  
the milk?

It's in the fridge



**What is the person doing?**



# What is the person doing?

In the conversation,



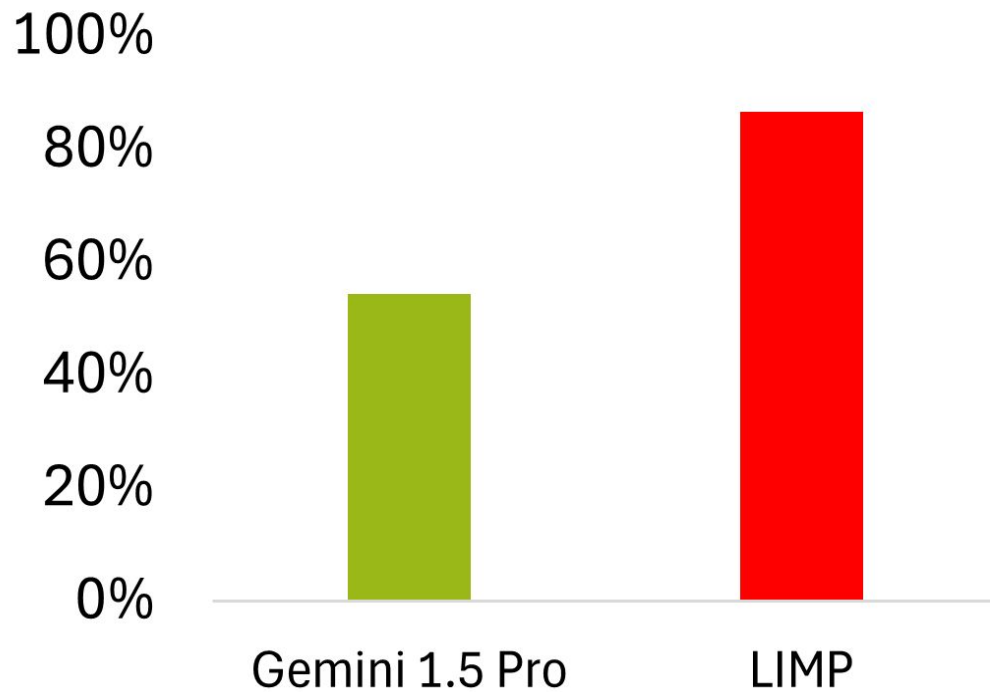
Where is  
the **milk?**

VLM might say: The person is  
grabbing a mug

LIMP: The person is grabbing **something**

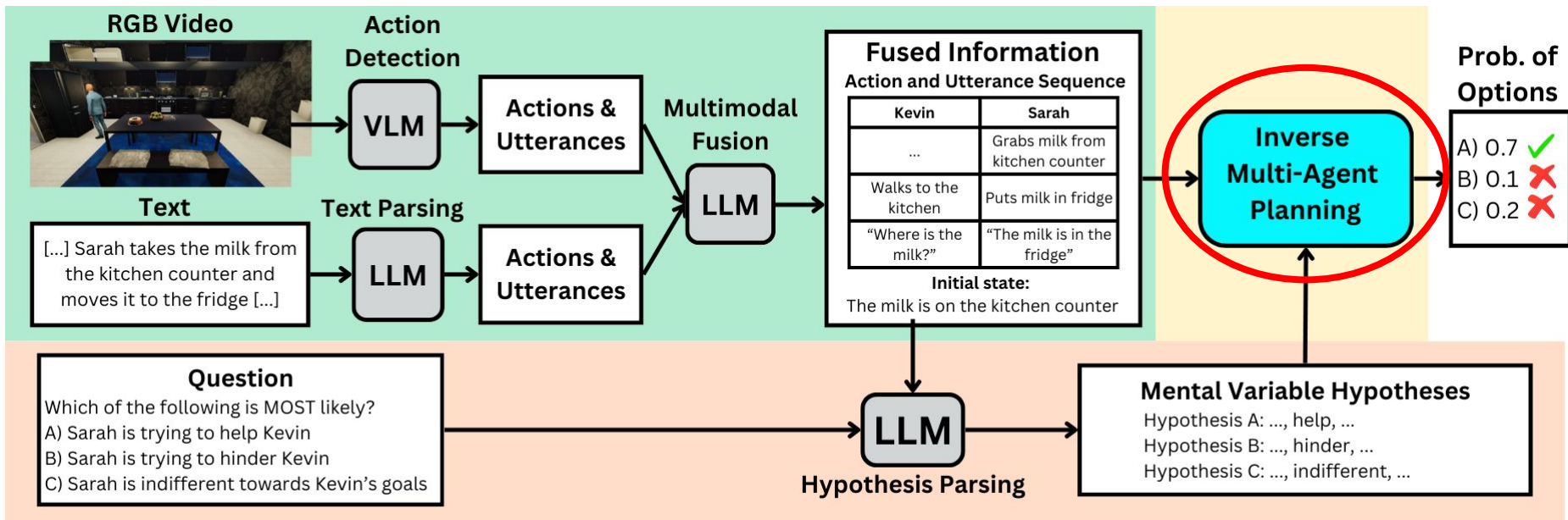
LIMP (Using Context):  
The person is grabbing  
**a bottle of milk**

## Action Extraction Accuracy





# LIMP (Language model-based Inverse Multi-modal Planning)





**Question:** If Jessica knows what is inside the cabinet in the bedroom, which of the following is **MOST** likely?

- A) Jessica is trying to help Kevin
- B) Jessica is trying to hinder Kevin
- C) Jessica is indifferent towards Kevin's goals

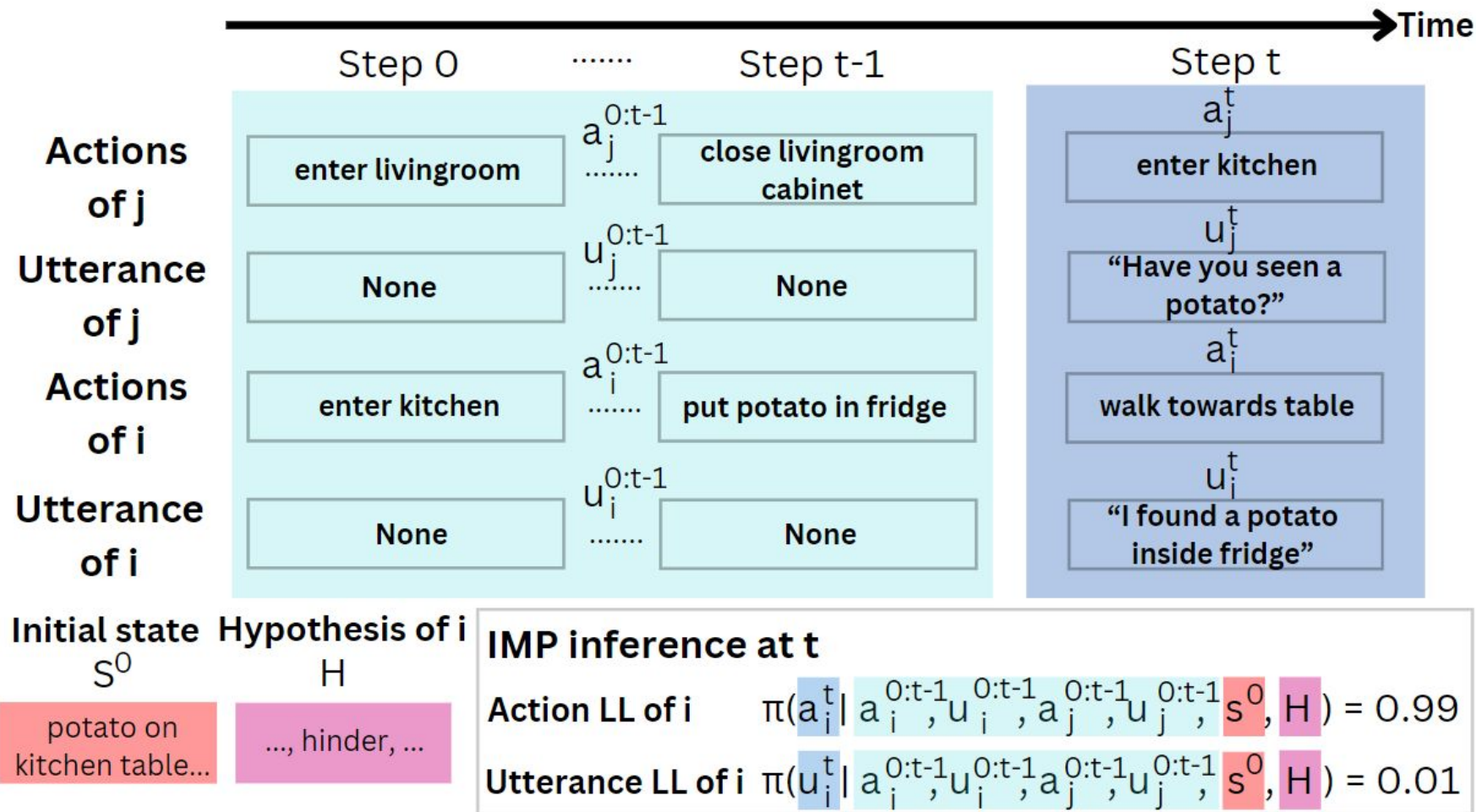
	Time →				
	Step 0	.....	Step 2	.....	Step 3
Actions of Kevin	enter kitchen	.....	open livingroom cabinet	.....	close livingroom cabinet
Utterance of Kevin	"Have you seen the magazine?"	.....	None	.....	None
Actions of Jessica	enter kitchen	.....	None (not shown)	.....	None (not shown)
Utterance of Jessica	"It's in the cabinet in the bedroom"	.....	None	.....	None

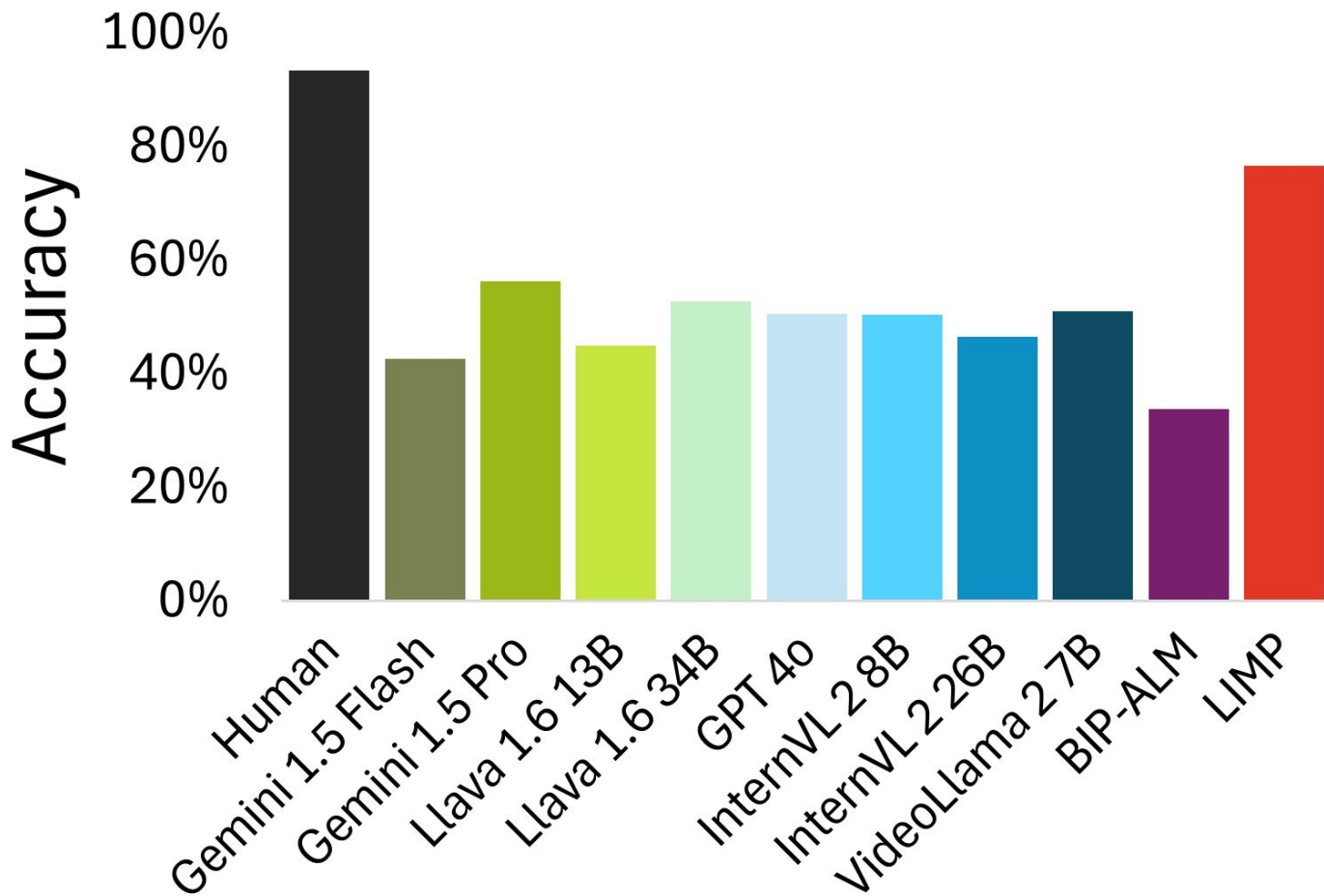
Suppose that Jessica knows the location of the magazine and want to help Kelvin ...

Is this true?

Where is the magazine? Not in the cabinet

Where does Jessica believe the magazine is? Somewhere other than the cabinet



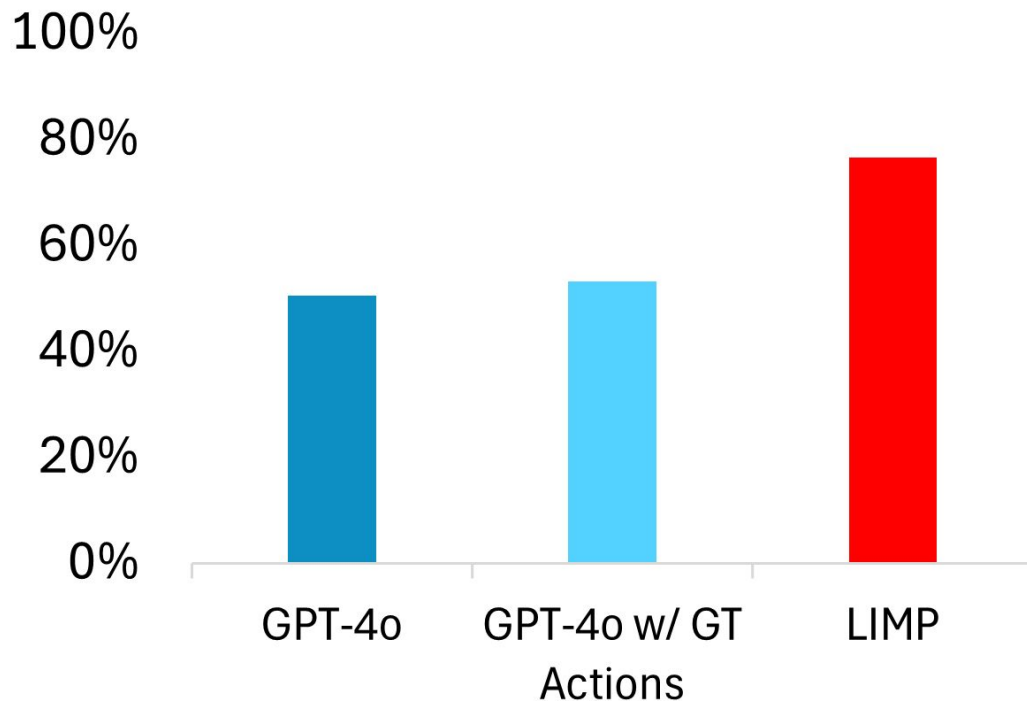




**Why do VLMs  
perform poorly?**

# Could it be due to poor action extraction?

Accuracy on MuMA-ToM

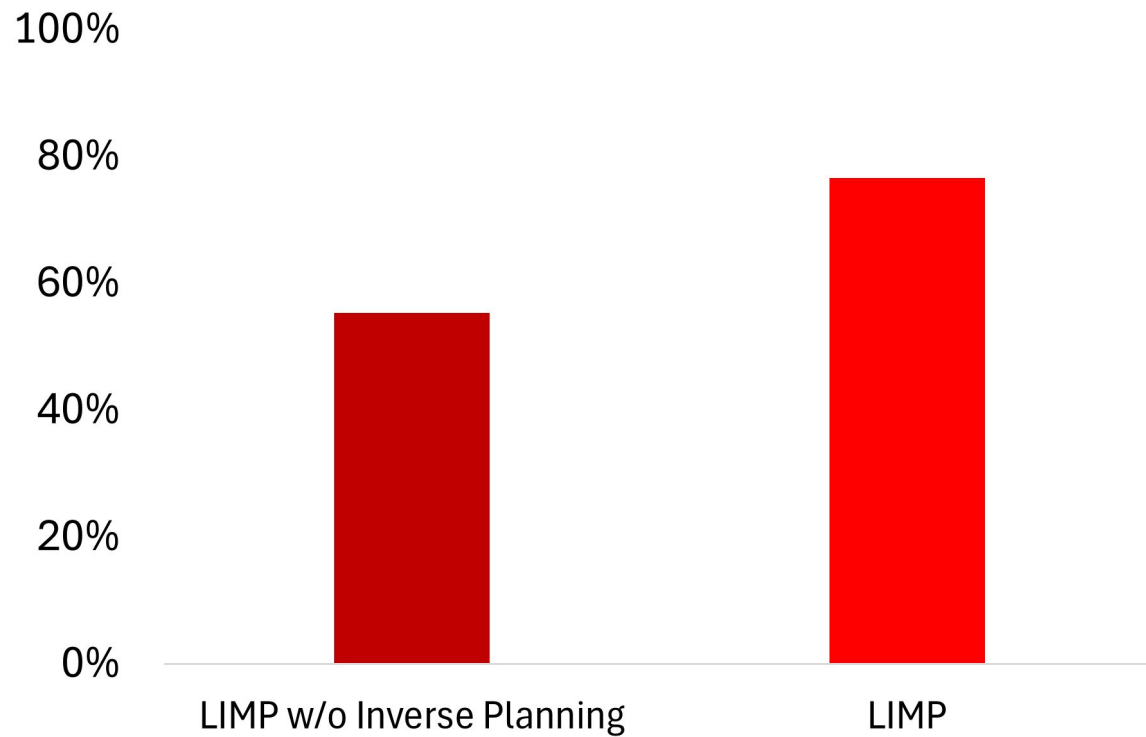






**Why does LIMP  
outperform the best  
VLMs?**

## Accuracy on MuMA-ToM



# Summary

- First **Multi-modal Multi-Agent** Theory of Mind Benchmark
- Existing LLMs perform very poorly on MuMA-ToM
- LIMP addresses problem through error correction and inverse planning



Code and Data