ToM-SSI: Evaluating Theory of Mind in Situated Social Interactions

cai
imprs-is
University of Stuttgart
Germany

Matteo Bortoletto, Constantin Ruhdorfer, Andreas Bulling
matteo.bortoletto@vis.uni-stuttgart.de

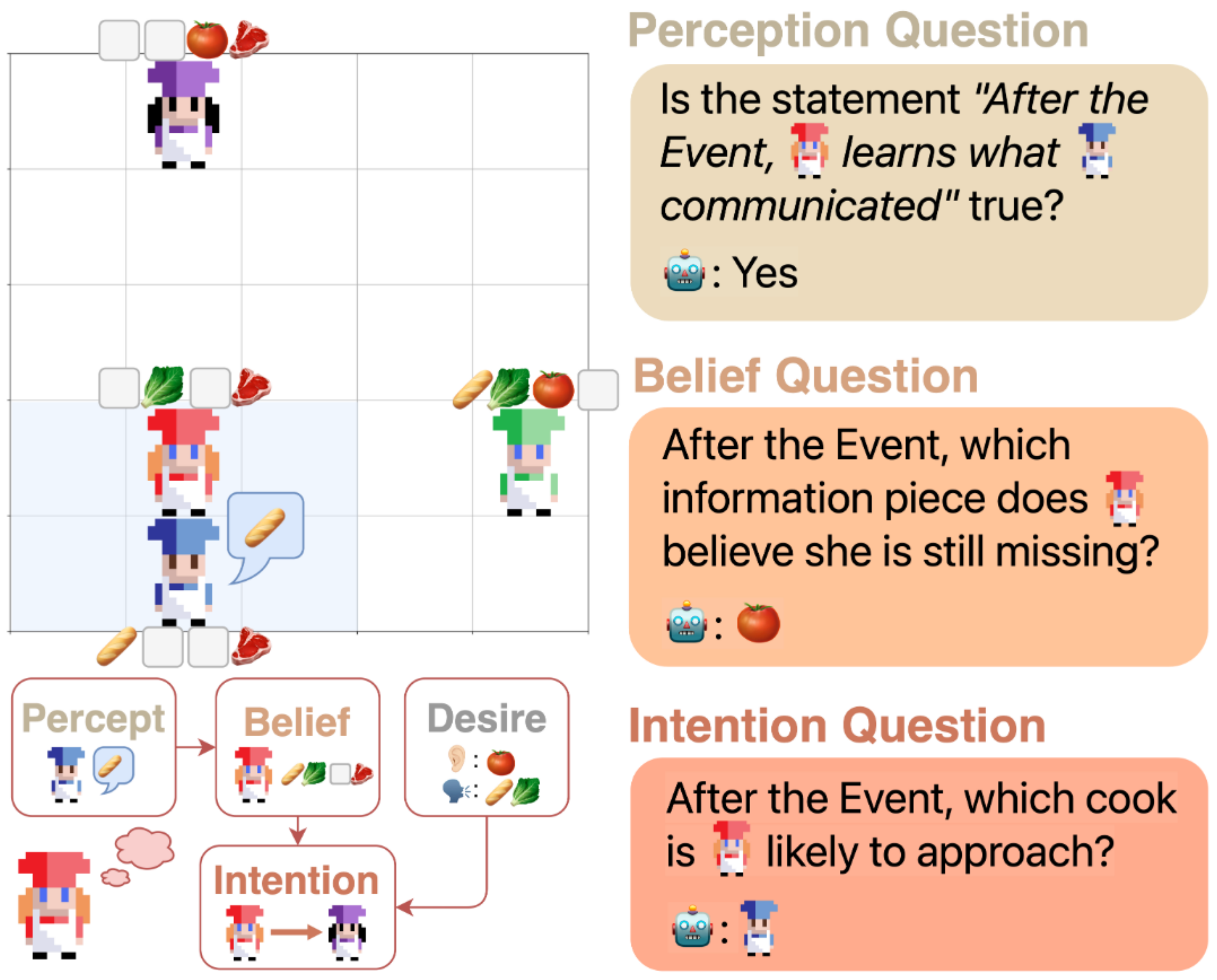
1. Motivation

- Theory of Mind (ToM) is the ability to attribute mental states to oneself and others [1].
- Most of current ToM benchmarks are text-based and/or variations of the Sally-Anne test [2]. They are also limited to one or two agents.
- ToM evaluations should be both physically and socially situated [3, 4].

2. Contributions

We present **ToM-SSI**, a multimodal benchmark that evaluates ToM abilities in situated social interactions:

- Formulated as a visual-text question answering task based on the Belief-Desire-Intention framework.
- Covers agent that move and communicate in a rich social environment with partial observability and constrained communication.
- Scenarios involve 3 or 4 agents, moving beyond dyadic interactions.
- It comprises 5 tasks covering cooperative, obstructive and mixed settings.



Perception Question

Is the statement *"After the Event, 🧑‍🍳 learns what 🧑‍🍳 communicated"* true?

🤖: Yes

Belief Question

After the Event, which information piece does 🧑‍🍳 believe she is still missing?

🤖: 🍅

Intention Question

After the Event, which cook is 🧑‍🍳 likely to approach?

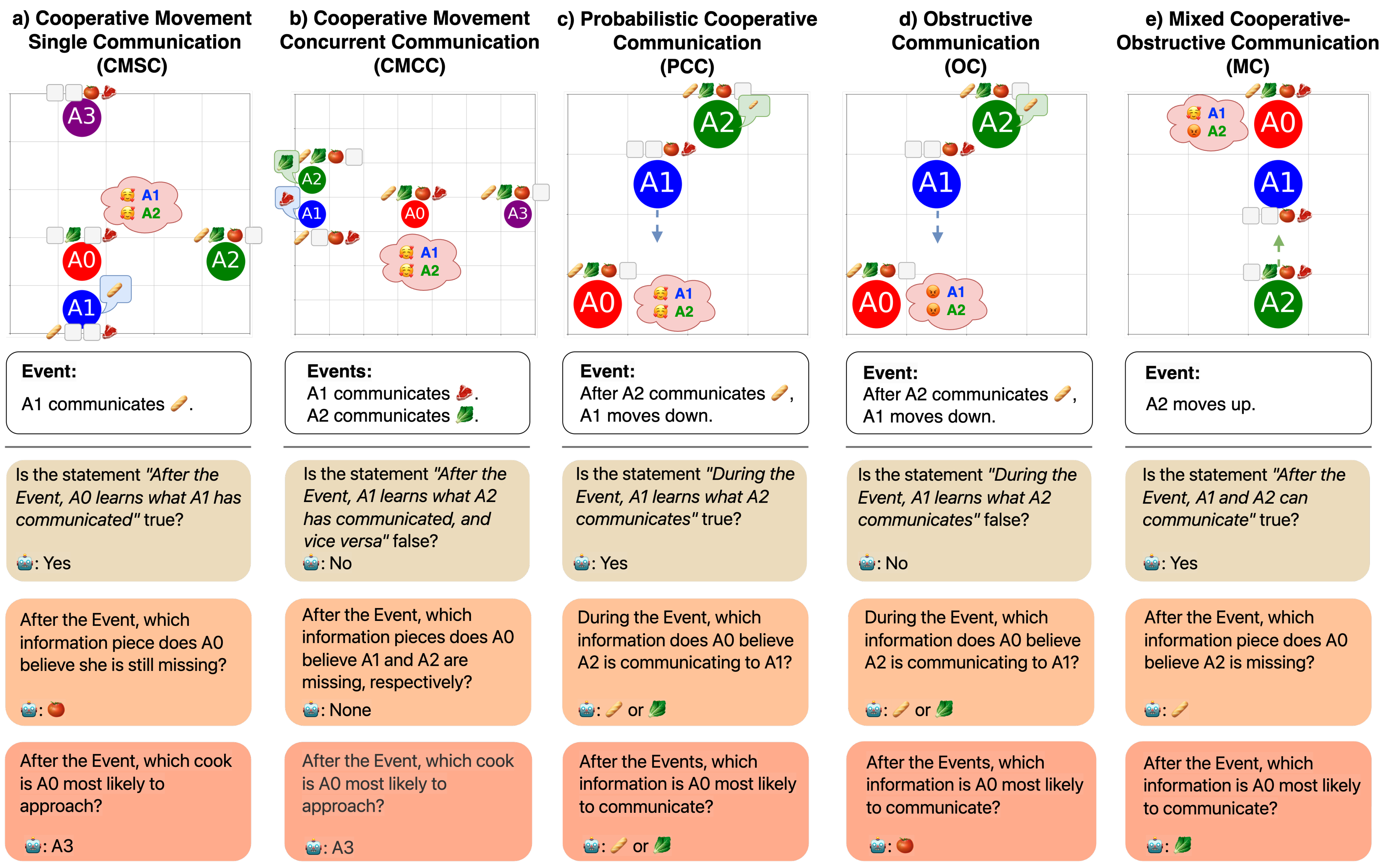
🤖: 🧑‍🍳

3. Tasks

Each sample is situated in a social context, e.g. *chefs in a kitchen preparing a dish*.

Events dictate the change in agent knowledge and the state of the environment. They involve movement and communication.

Five tasks, each including 3 question types: percept, belief, intent. Desire is fixed by the agent attitude (collaborative, obstructive, mixed).



a) Cooperative Movement Single Communication (CMSC)

Event: A1 communicates 🥖.

Is the statement *"After the Event, A0 learns what A1 has communicated"* true?
🤖: Yes

After the Event, which information piece does A0 believe she is still missing?
🤖: 🍅

After the Event, which cook is A0 most likely to approach?
🤖: A3

b) Cooperative Movement Concurrent Communication (CMCC)

Events: A1 communicates 🥩. A2 communicates 🥬.

Is the statement *"After the Event, A1 learns what A2 has communicated, and vice versa"* false?
🤖: No

After the Event, which information pieces does A0 believe A1 and A2 are missing, respectively?
🤖: None

After the Event, which cook is A0 most likely to approach?
🤖: A3

c) Probabilistic Cooperative Communication (PCC)

Event: After A2 communicates 🥖, A1 moves down.

Is the statement *"During the Event, A1 learns what A2 communicates"* true?
🤖: Yes

During the Event, which information does A0 believe A2 is communicating to A1?
🤖: 🥖 or 🥬

After the Events, which information is A0 most likely to communicate?
🤖: 🥖 or 🥬

d) Obstructive Communication (OC)

Event: After A2 communicates 🥖, A1 moves down.

Is the statement *"During the Event, A1 learns what A2 communicates"* false?
🤖: No

During the Event, which information does A0 believe A2 is communicating to A1?
🤖: 🥖 or 🥬

After the Events, which information is A0 most likely to communicate?
🤖: 🍅

e) Mixed Cooperative-Obstructive Communication (MC)

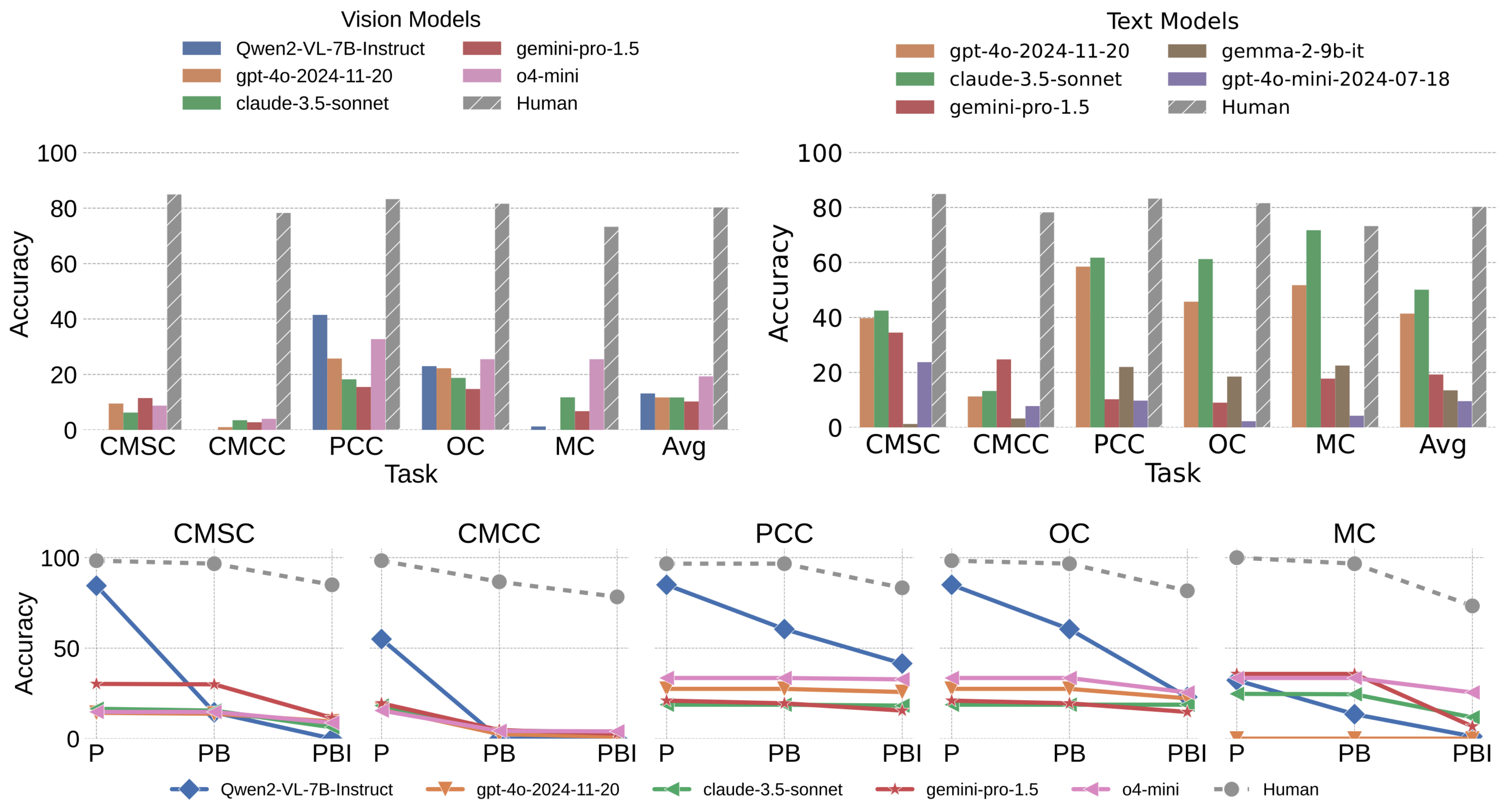
Event: A2 moves up.

Is the statement *"After the Event, A1 and A2 can communicate"* true?
🤖: Yes

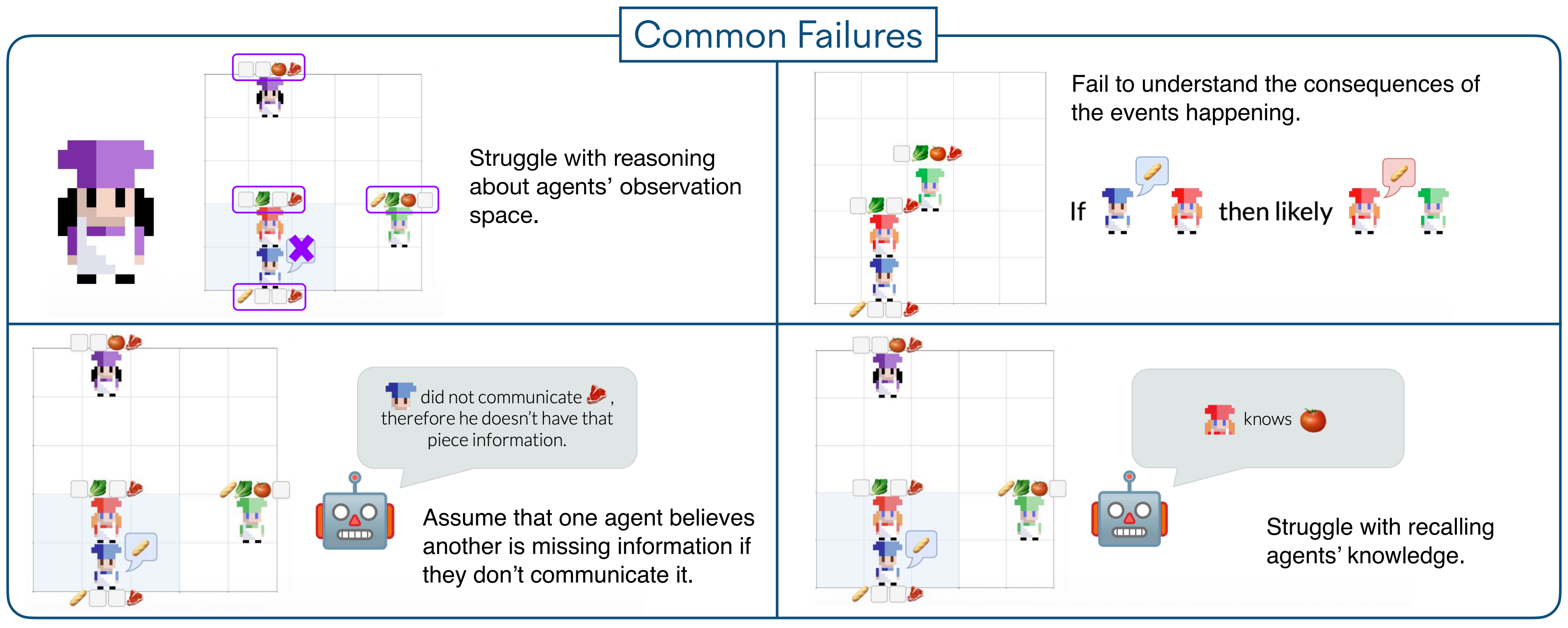
After the Event, which information piece does A0 believe A2 is missing?
🤖: 🥖

After the Event, which information is A0 most likely to communicate?
🤖: 🥬

4. Experiments



5. Key Findings



Common Failures

Struggle with reasoning about agents' observation space.

Fail to understand the consequences of the events happening.

If 🧑‍🍳 🧑‍🍳 then likely 🧑‍🍳 🧑‍🍳

🧑‍🍳 did not communicate 🥩, therefore he doesn't have that piece information.

Assume that one agent believes another is missing information if they don't communicate it.

🧑‍🍳 knows 🍅

Struggle with recalling agents' knowledge.

- Performance gap between models and humans.
- Models struggle with the critical steps for ToM reasoning.
- Challenges: modelling other agents' perception, multi-agent communication, and mixed social interactions.
- VLMs are not able yet to consistently combine textual and visual information.

References

[1] Premack, David, and Guy Woodruff. "Does the chimpanzee have a theory of mind?." *Behavioral and brain sciences* 1.4 (1978): 515-526.

[2] Gandhi, Kanishk, et al. "Understanding social reasoning in language models with language models." *NeurIPS* 2024.

[3] Ma, Ziqiao, et al. "Towards A Holistic Landscape of Situated Theory of Mind in Large Language Models." *Findings of EMNLP* 2023.

[4] Bortoletto, Matteo, et al. "Limits of Theory of Mind Modelling in Dialogue-Based Collaborative Plan Acquisition." *ACL* 2024.

30th ANNIVERSARY EMNLP 2025
Suzhou, China November 4-9
中国苏州 11月4日-9日