NEURIPS 2026

Mental-Models for Multi-Agent Systems

Hanan Gani · Lulu Shao · Manmohan Chandraker

University of California, San Diego

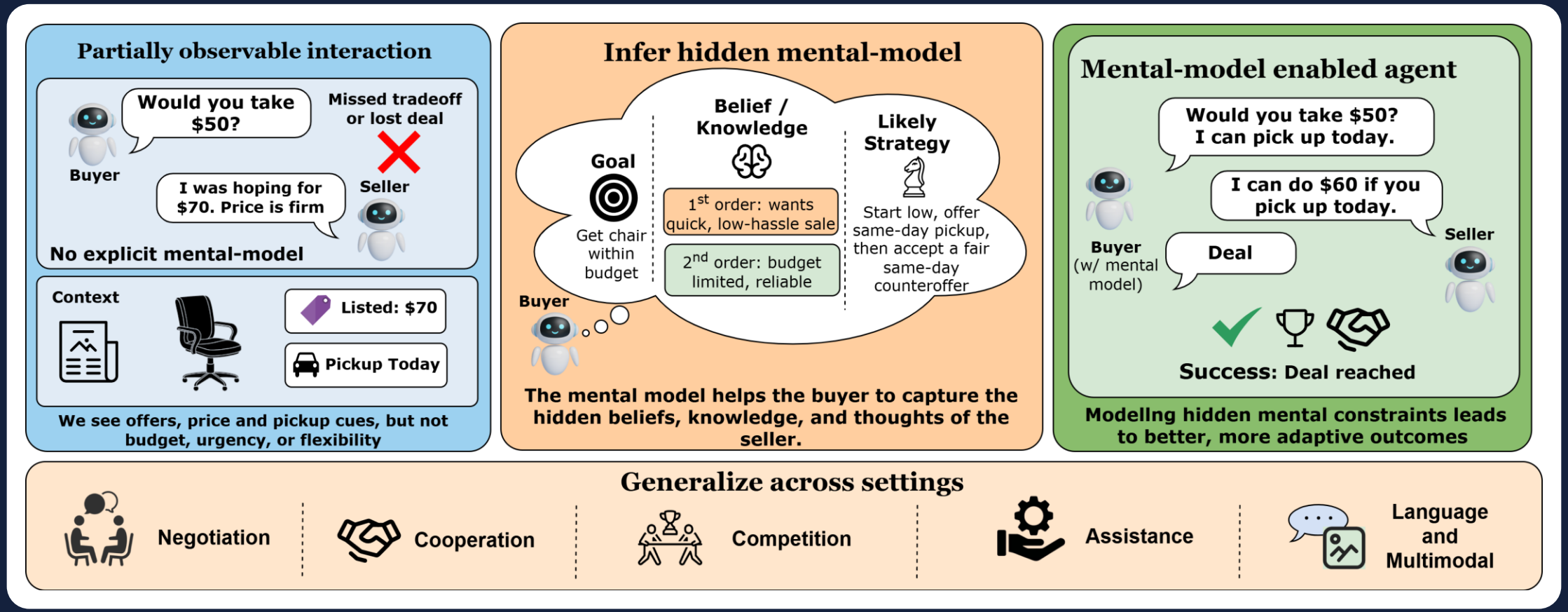


UC San Diego



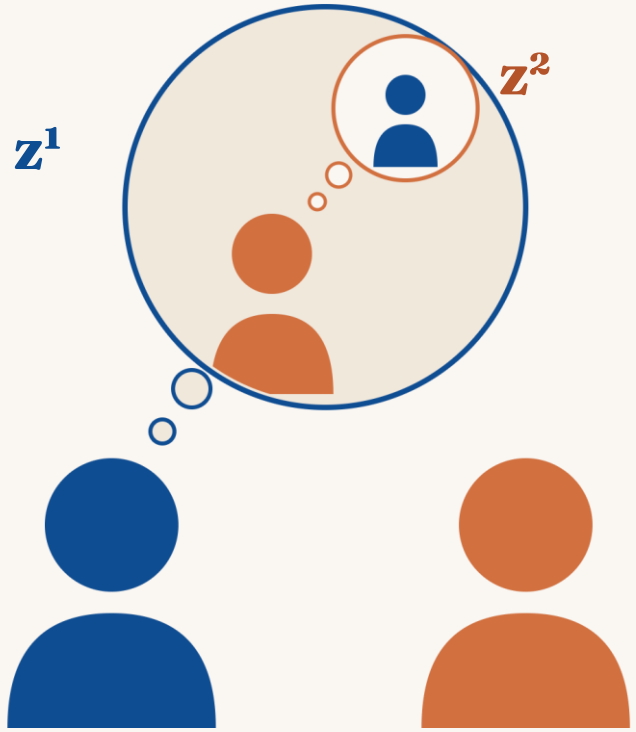
THE PROBLEM

Agents hear words, not minds



THE IDEA

Give the agent a model of its partner's mind



z^1 Their mind

What they believe, intend and think

Belief · Intent · Thought

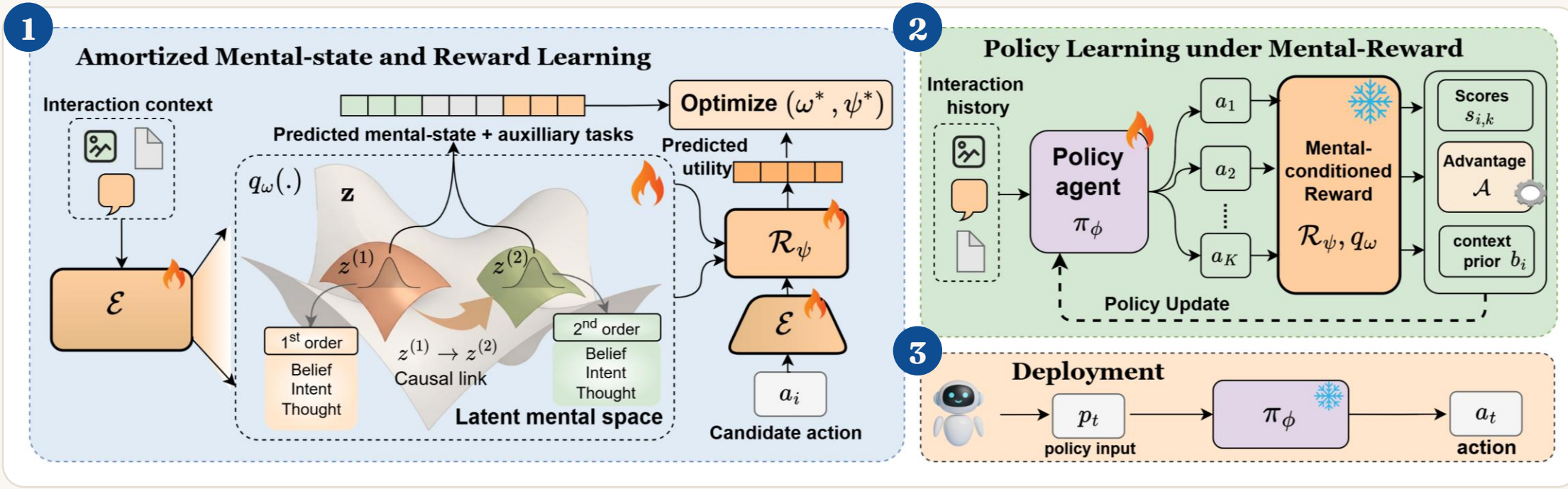
z^2 Their view of me

What they believe about the agent

Both inferred from the conversation, with uncertainty.

THE METHOD

From mental model to policy



1 Learn the mind

Infer z¹ and z², and learn a reward on them

2 Learn to act

Frozen reward + GRPO

3 Act alone

Only the policy runs

RESULTS

One recipe, gains everywhere

+11.5%

SOTOPIA

open-ended social interaction

+26.7

BigToM

points of theory-of-mind accuracy

+3.9%

MMRole

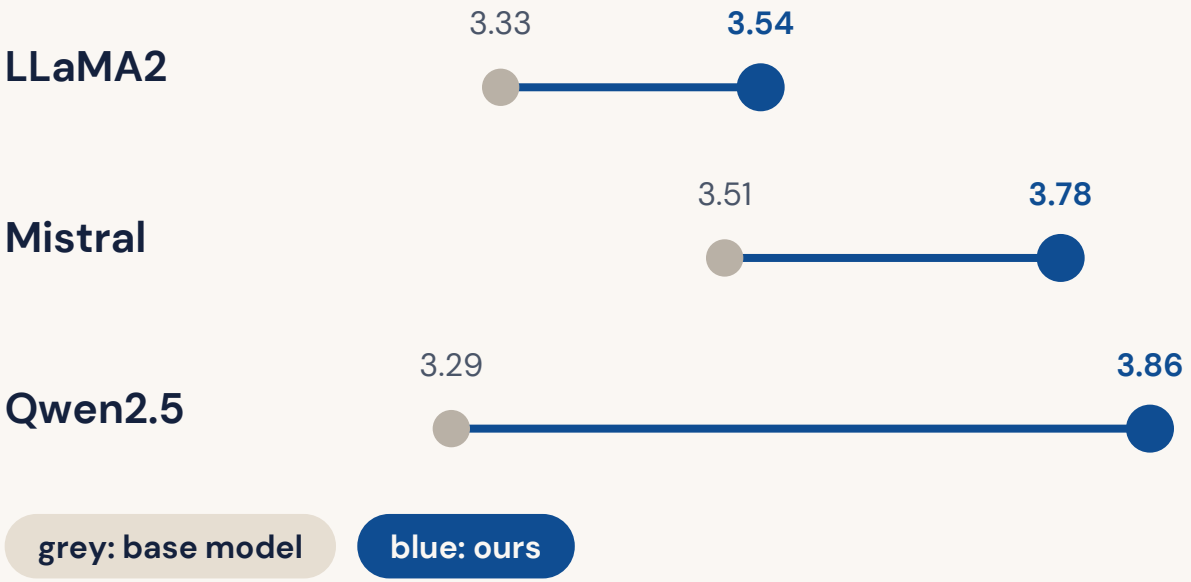
vision-language role-play, unseen roles

RESULTS

Small models, strong results

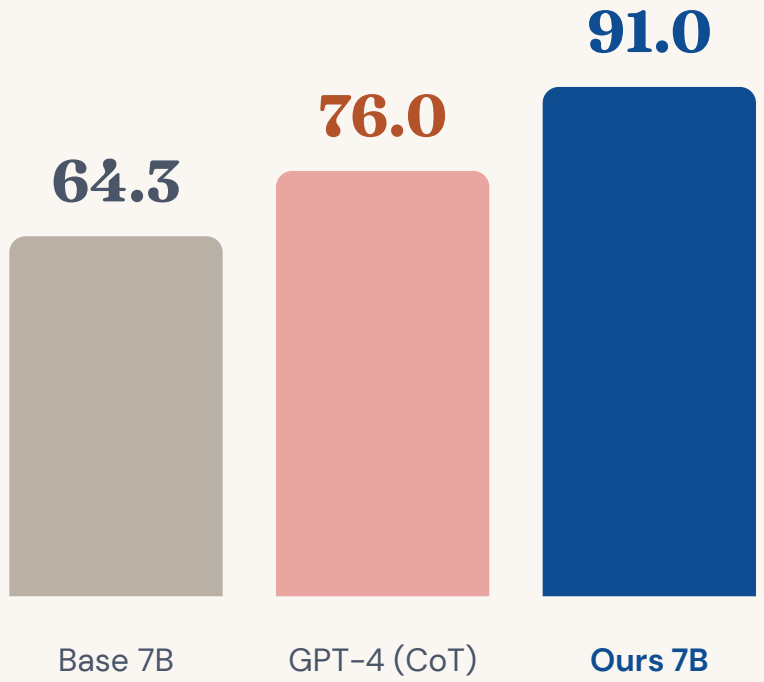
SOTOPIA

Overall score of 7B agents, base vs. ours



BigToM

Mean paired accuracy (%), six settings



ZERO-SHOT TRANSFER

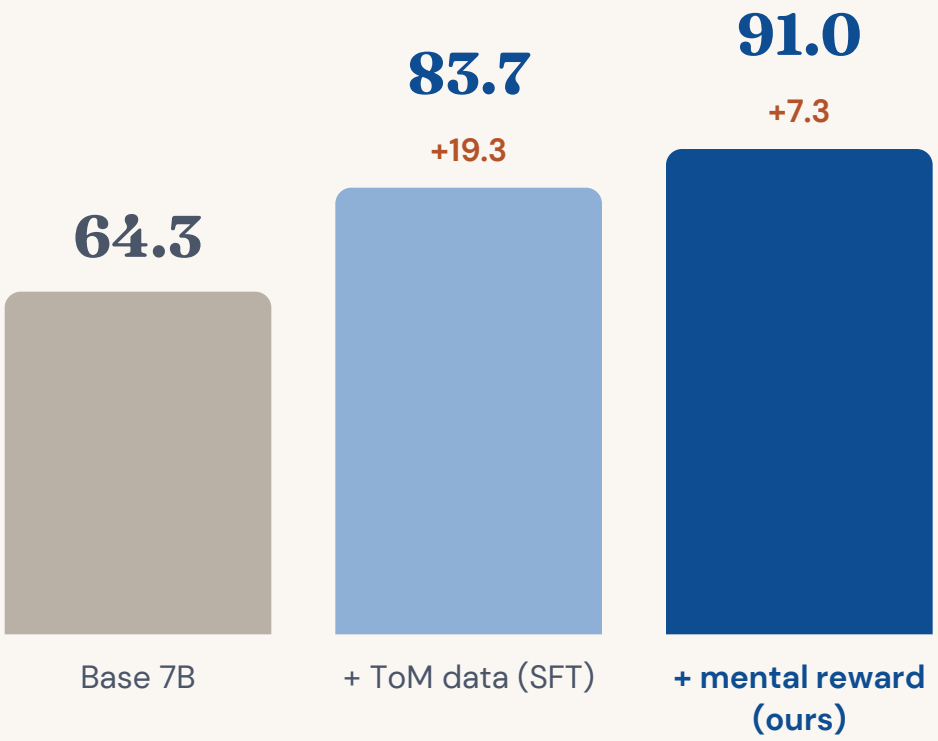
Skills that travel

SOTOPIA	→ Craigslist-Bargain	91% → 94%	deal rate in price negotiations
BigToM	→ ToMi	55.2 → 60.5	joint belief-tracking accuracy
BigToM	→ FANToM	49.7 → 52.2	second-order belief, multi-party chats

ANALYSIS

Where the gains come from

BigToM mean paired accuracy (%)



LABEL EFFICIENCY · SOTOPIA

LABELS USED **25%**

GAIN RECOVERED **63%**

A quarter of the labels, most of the gain.

INSIDE THE MENTAL MODEL

The reward listens to z

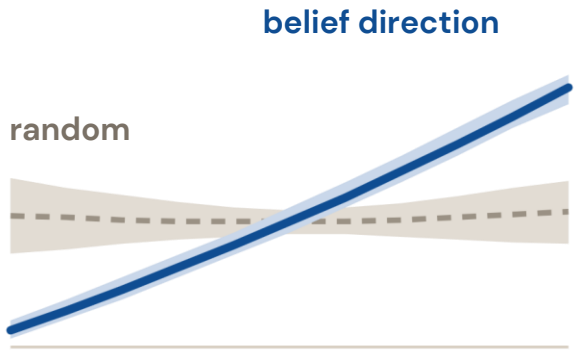
Scramble z

Mental-model z **~0.08**

Compression VAE **~0**

Rise in reward-prediction error when z is scrambled

Steer z



Frozen reward as z moves along a direction

Shuffle its labels

B·I·T labels (ours) **25.06**

Shuffled labels **12.53**

Median reward margin, preferred vs. rejected reply

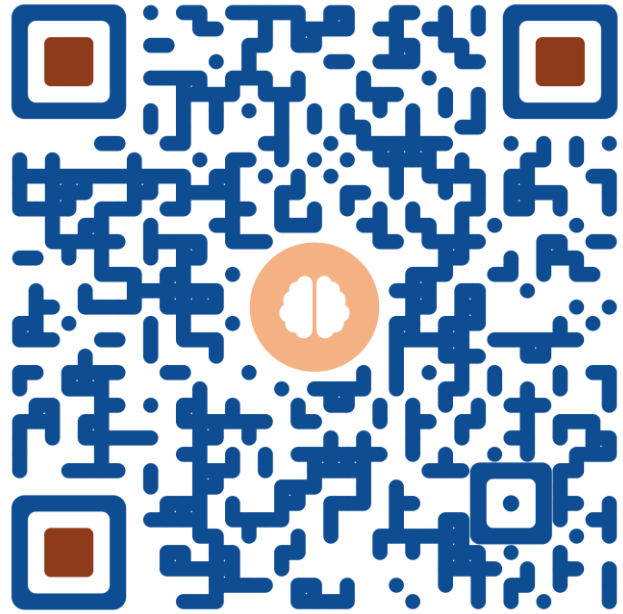
TAKEAWAY

Model the mind, then act.

One recipe for social, theory-of-mind and vision-language agents — at no extra cost at test time.

hananshafi.github.io/Mental-Models

arXiv 2610.12453 · github.com/hananshafi/Mental-Models



Project page

paper · code · data