UCF | Institute of Artificial Intelligence

UNIVERSITY OF CENTRAL FLORIDA

OmniToM: Benchmarking Theory of Mind in LLMs via Explicit Belief Modeling

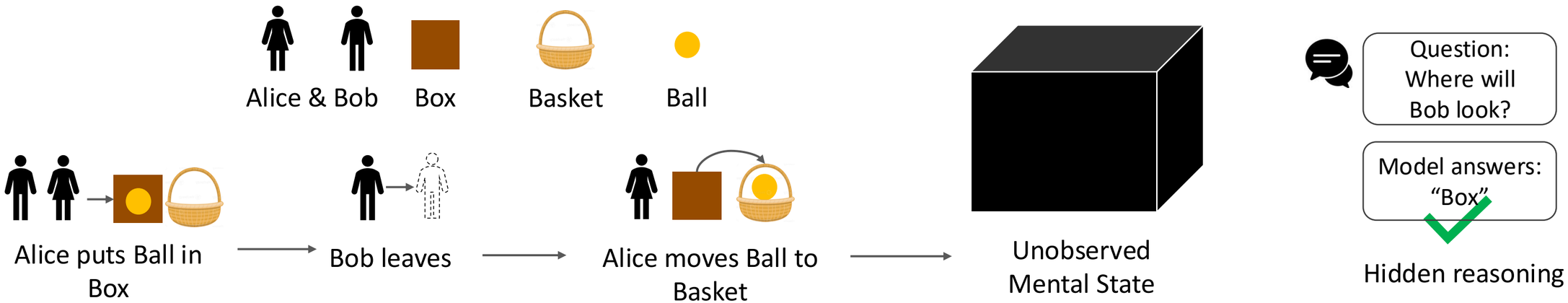
NeurIPS 2026

Adam Bawatneh · Sagar Sapkota · Amrit Singh Bedi · Santu Karmaker · Mubarak Shah

Motivation and Problem

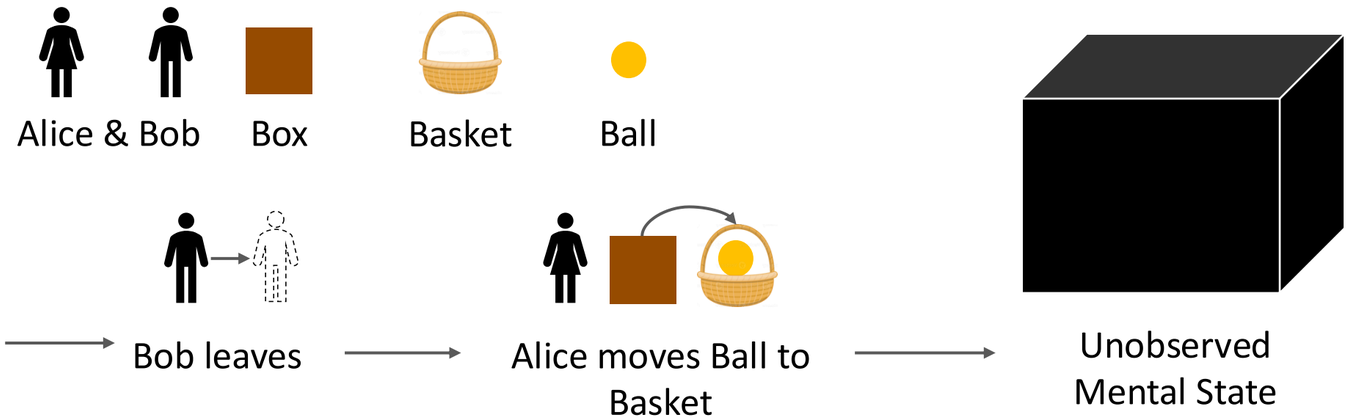
- Theory-of-Mind (ToM) is the ability to attribute mental states: beliefs, intentions, etc.
- Supports social reasoning and safe interaction.
- Endpoint QA can miss errors in models' representations of mental states.

Endpoint Question Answering (Q/A)



Motivation and Problem

Endpoint Question Answering (Q/A)

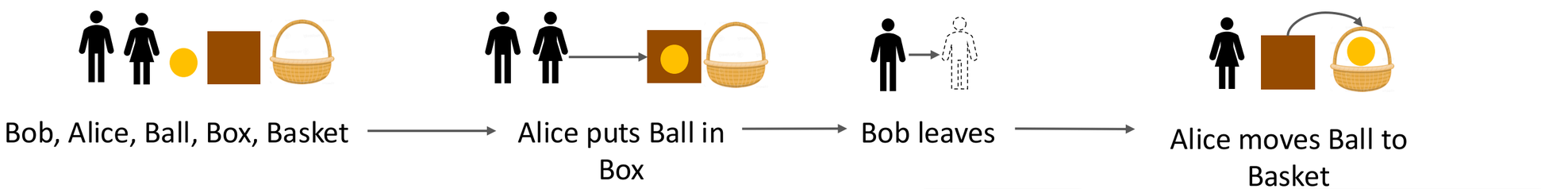


Question: Where will Bob look?

Model answers: "Box"

Hidden reasoning

Mental-State Representation via Beliefs



Alice believes: -	Bob believes: -	Alice believes: Ball in Box	Bob believes: Ball in Box	Alice believes: Bob leaves	Bob believes: Ball in Box	Alice believes: Ball in Basket	Bob believes: Alice believes Ball in **Basket**

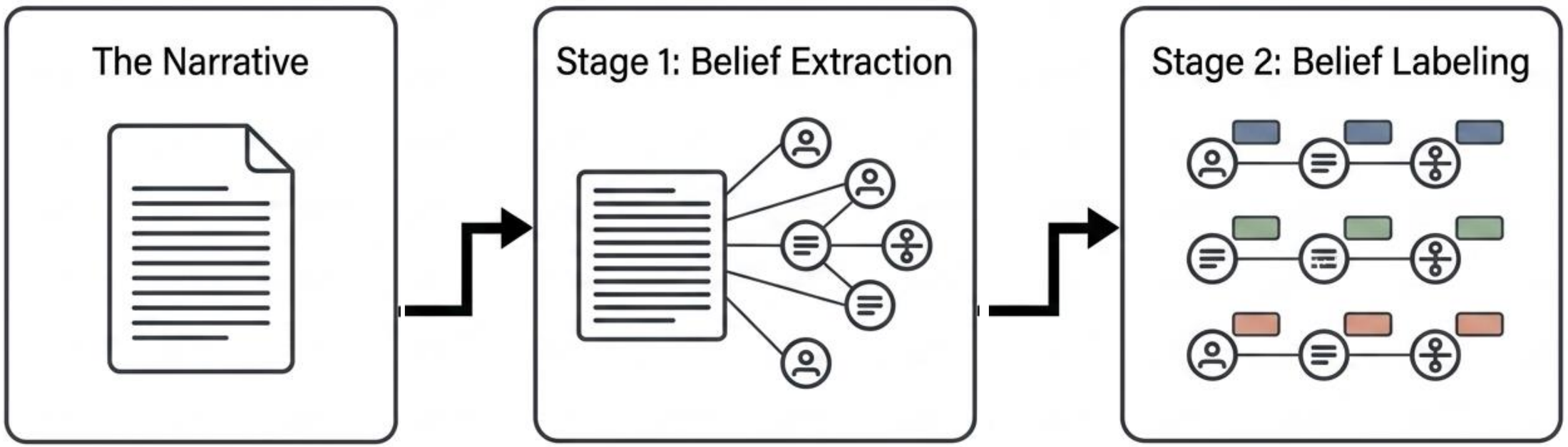
Question: Where will Bob look?

Model answers: "Box"

Endpoint answer is correct **OmniToM detects error.** Bob should believe that Alice believes the ball is in the box.

A Two-Stage Evaluation Framework

OmniToM splits the extraction of the beliefs from their fine-grained cognitive labeling.

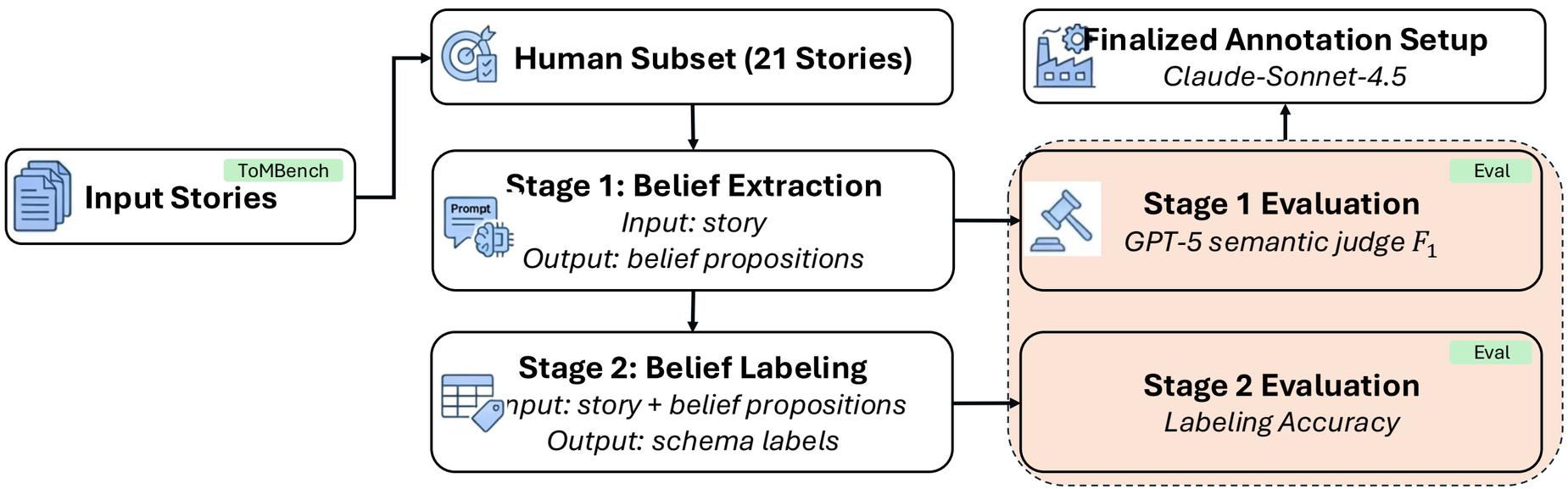


Benchmark Construction

895
Social Reasoning Stories

22,343
Labeled Belief Propositions

156,401
Schema Labels



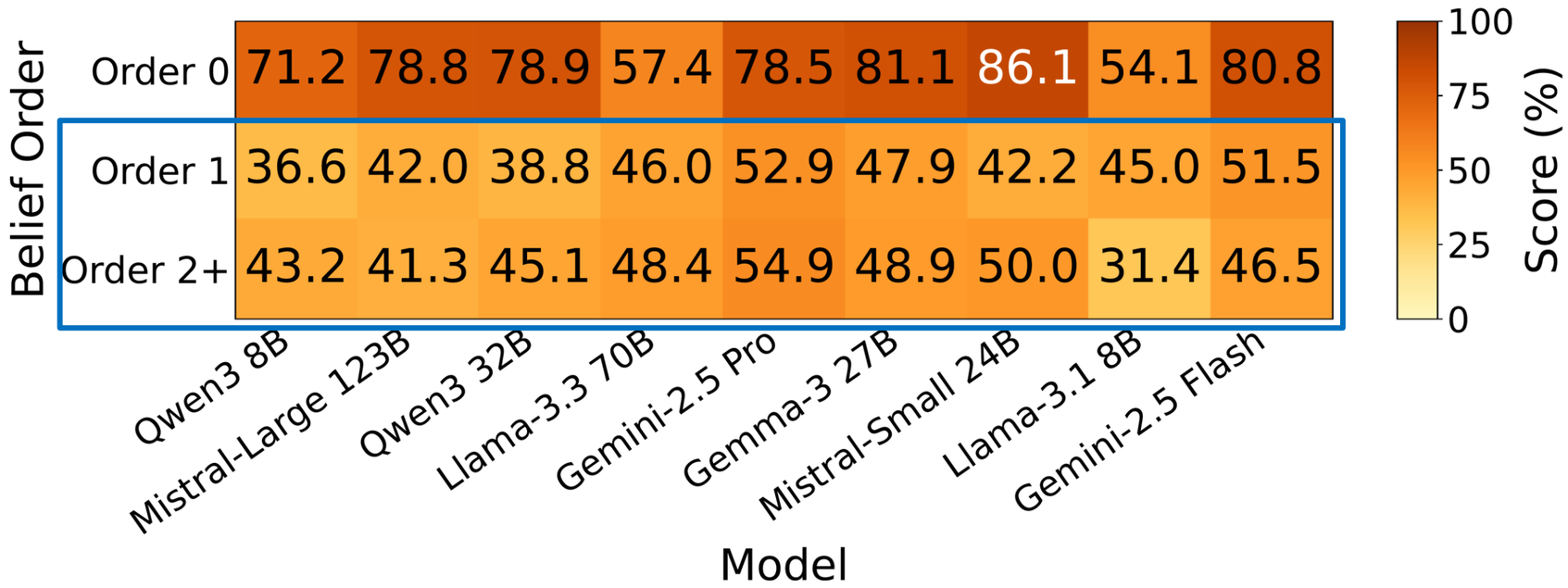
Main Results

- Model rankings differ between ToMBench Endpoint QA and OmniToM's explicit belief evaluation.
- Gemma-3 27B ranks second on Stage 1 extraction but ninth on Endpoint QA.
- Stage 2 errors in Knowledge Access and Representation point to an information-tracking bottleneck.

	Stage 1	Endpoint QA	Stage 2: Belief-Labeling Accuracy							
Model	F_1 (95% CI)	Accuracy (95% CI)	Order	Status	Access	Repr.	CType	Source	Context	Overall
Closed-source models										
Gemini-2.5 Pro	**61.63 ± 0.76**	**80.67 ± 2.15**	96.23	86.50	66.94	**89.01**	**86.35**	**86.70**	87.07	85.54
Gemini-2.5 Flash	55.75 ± 1.10	77.30 ± 2.32	95.56	84.97	71.34	87.58	85.97	84.10	92.14	**85.95**
GPT-5	–	–	95.18	82.72	66.85	83.42	79.96	83.02	88.83	82.85
Open-weight models										
Gemma-3 27B	58.03 ± 0.98	53.95 ± 2.53	96.56	82.44	71.57	54.33	73.50	78.72	92.07	78.46
Mistral-Small 24B	56.46 ± 0.94	71.08 ± 2.45	95.13	82.22	**74.59**	62.79	76.01	84.82	91.90	81.06
Mistral-Large 123B	53.53 ± 0.92	73.50 ± 2.42	**97.25**	**86.53**	74.14	72.87	82.83	86.32	**92.97**	84.70
Qwen3 32B	51.63 ± 0.94	69.96 ± 2.44	96.42	82.43	73.91	62.45	71.27	76.84	90.81	79.16
Llama-3.3 70B	47.12 ± 0.99	71.28 ± 2.45	92.74	83.55	67.41	72.43	72.35	76.71	91.69	79.55
Qwen3 8B	42.70 ± 1.18	63.31 ± 2.46	73.38	67.17	57.94	63.77	51.43	61.49	74.62	64.26
Llama-3.1 8B	37.21 ± 1.18	60.79 ± 2.58	71.90	65.59	56.13	64.40	48.63	55.18	76.81	62.66

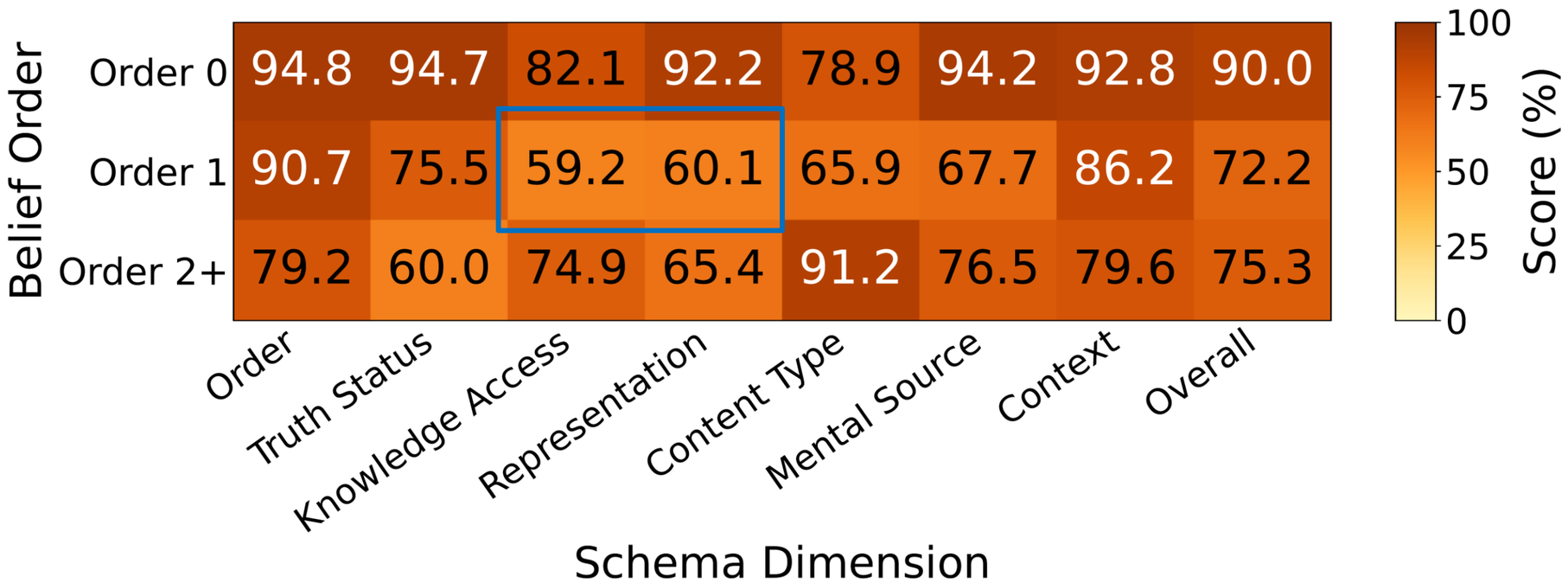
Stage 1: Actor-specific beliefs are an extraction bottleneck

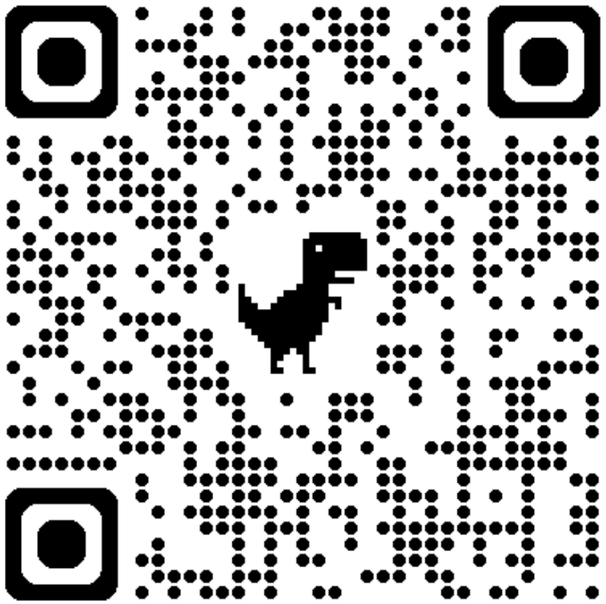
- Models extract narrated facts (Order 0) more reliably than actor-specific (Order 1+) beliefs.
- The gap appears when they must track what each person knows or believes.



Stage 2: Information Tracking Bottleneck

- Stage 2 accuracy is weakest at Order 1 for Knowledge Access and Representation.
- Since Order 2+ accuracy is higher overall, the errors do not increase with belief depth.





Project Page