國立臺灣大學
National Taiwan University

擺脫學習偷吃步：
打造穩健的語言模型心智推理後訓練方法

羅紹元

助理教授、玉山青年學者

國立臺灣大學資訊工程學系

12/07/2025

羅紹元 Shao-Yuan Lo

國立臺灣大學
National Taiwan University

HRI
Honda Research Institute US

amazon

JOHNS HOPKINS
UNIVERSITY

國立交通大學
National Chiao Tung University

- **國立臺灣大學**資訊工程學系 助理教授（2025 至今）
- **本田美國研究院** AI科學家（2023 – 2025）
- **亞馬遜公司** 實習研究員（2021、2022 暑假）

- **約翰霍普金斯大學**電機與電腦工程 博士（2019 – 2023）
- **國立交通大學**電子研究所 碩士（2017 – 2019）
- **國立交通大學**電機資訊學士班 學士（2013 – 2017）

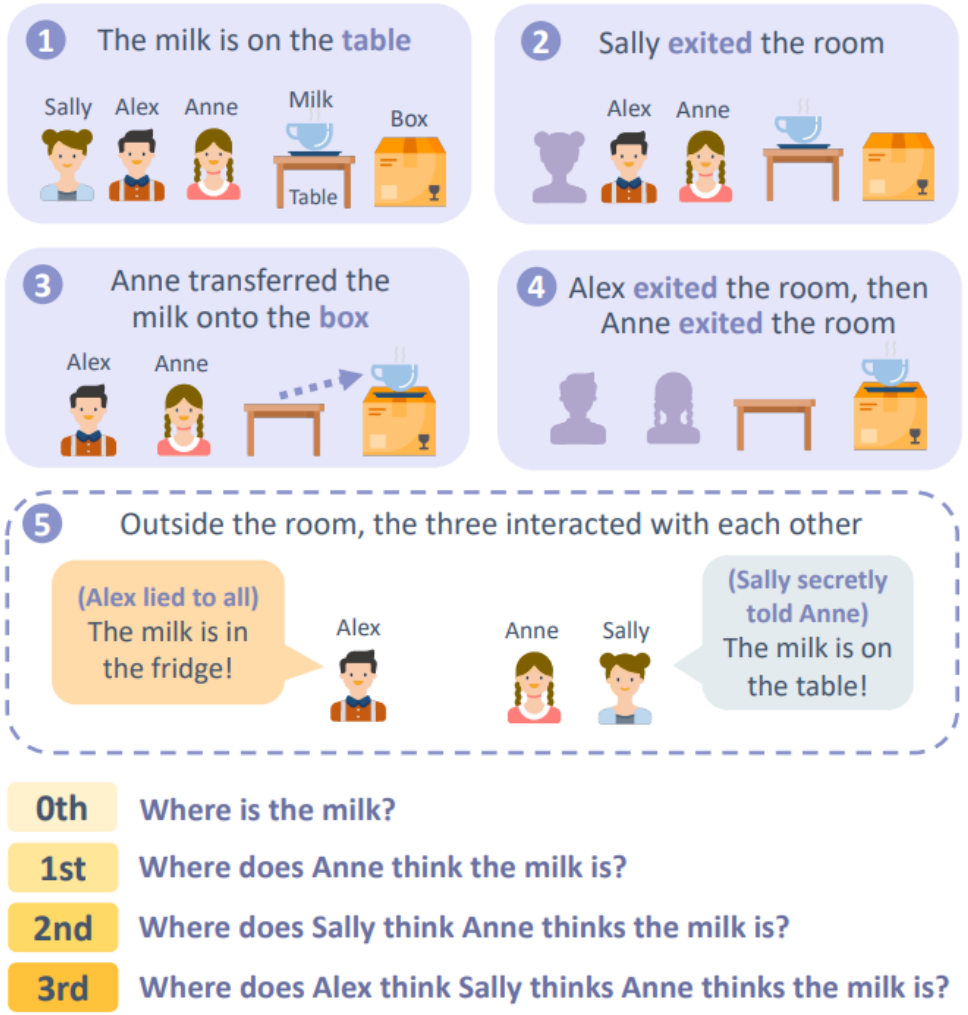
Outline

- What is Theory of Mind (ToM)?
- What is shortcut learning?
- Shortcut learning in ToM
- Robust evaluation of ToM post-training

What is Theory of Mind?

- ToM is the ability to **understand other people's mental states**, such as thoughts, emotions, intentions, and beliefs
- **Machine ToM** aims to replicate this human's innate ability in AI agents

[He et al. EMNLP-Findings'23]



ToM Promotes Safe Human–AI Collaboration

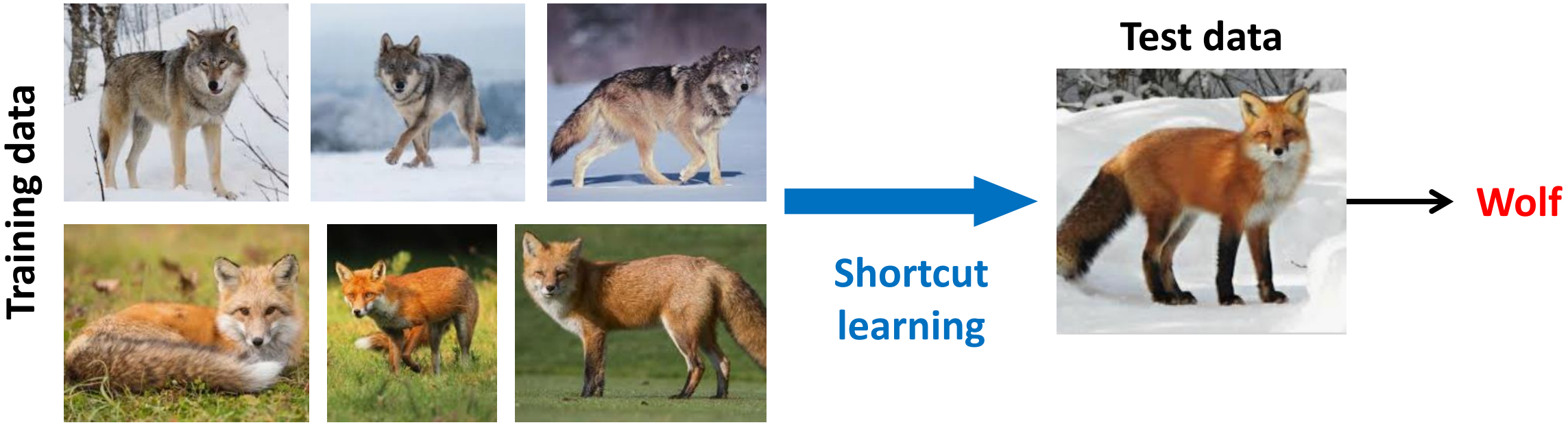
- Infer user's mental states, such as thoughts, intentions, and beliefs
- Track perspectives (what the human knows vs. doesn't know)
- Predict next actions and anticipate needs

- → More aligned assistance
- → Fewer misunderstandings
- → Safer interactions

Image source: https://medium.com/@muhammadahmedsuhail007/embracing-human-ai-collaboration-for-the-future-of-work-52cbc62a0c8a

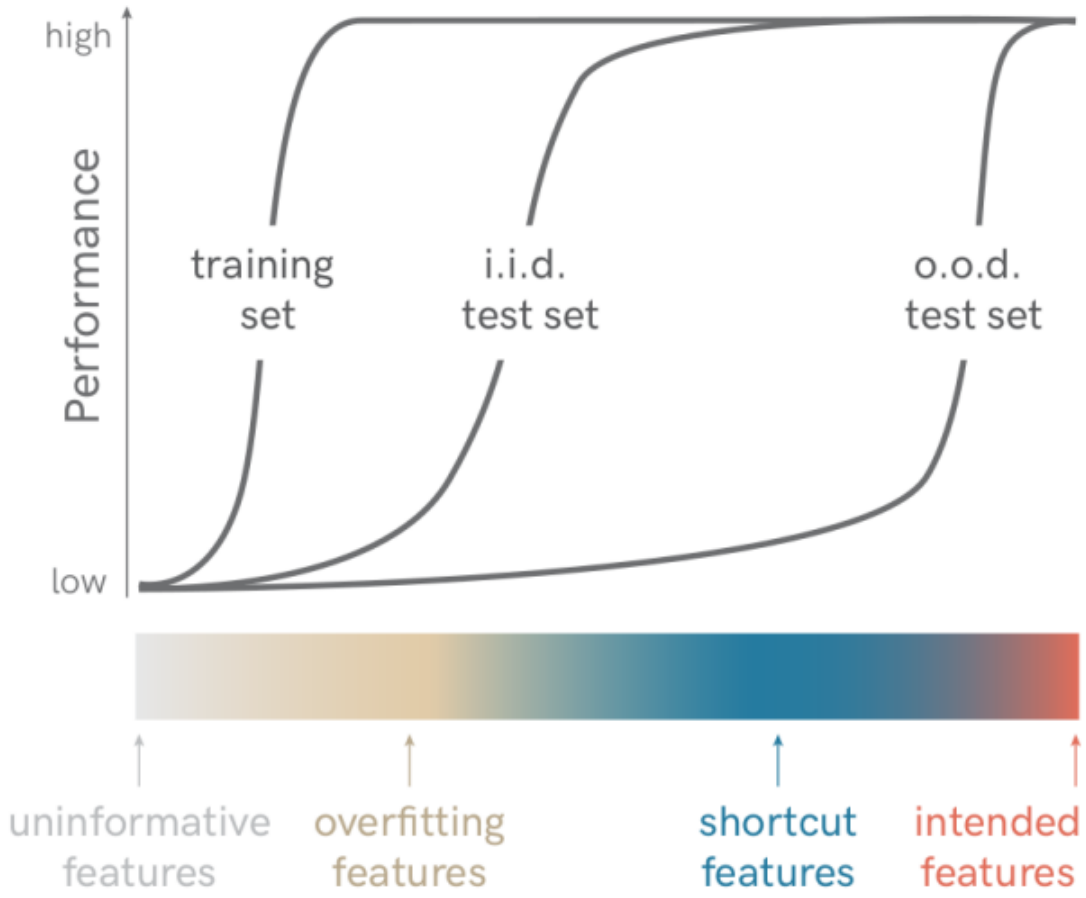
What is Shortcut Learning?

- A model achieves high accuracy by **exploiting easy, spurious patterns** in the data instead of learning the **true underlying concepts**
- E.g., a wolf–fox classifier may just learn “snow = wolf” if all the wolf training images include snow



What is Shortcut Learning?

- How a model relying on different types of features performs across different test sets
- Some features fit the training data; fewer generalize to i.i.d. tests
- Among those, shortcut features fail under OOD shifts, while only the intended features truly generalizes



[Geirhos et al. Nature Machine Intelligence'20]

What We Observed in ToM?

- State-of-the-art LLMs do not perform well on ToM tasks (around 70% accuracy)

Model	4th-order ToM	Hi-ToM	ToMi	ExploreToM (Raw)	ExploreToM (Infilled)
GPT-4o	46.17%	69.00%	61.96%	67.35%	62.48%
GPT-4o-mini	30.50%	58.50%	**70.64%**	**69.32%**	**66.04%**
DeepSeek-v3	**58.67%**	**70.17%**	57.17%	65.01%	64.73%

- ToM's property: long, complicated, contains causal relationships
- Intuition: use **reasoning**
- → Can we improve LLMs' ToM via **RL post-training**?

[Lu et al. arXiv'25]

What We Observed in ToM?

- A new study on **ToM post-training** appeared on arXiv in **May 2025**
- **Their key takeaway**: SFT achieves competitive performance with RL on current ToM benchmarks for ToM post-training

---

Do Theory of Mind Benchmarks Need Explicit Human-like Reasoning in Language Models?

Yi-Long Lu,*† **Chunhui Zhang**,*† **Jiajun Song**, **Lifeng Fan**, **Wei Wang**†
State Key Laboratory of General Artificial Intelligence, BIGAI, Beijing, China
luyilong@pku.edu.cn, {zhangchunhui, songjiajun, lifengfan, wangwei}@bigai.ai

arXiv, May 2025

What We Observed in ToM?

- **Their key takeaway**:
  SFT achieves competitive performance with RL on current ToM benchmarks for ToM post-training

- Interesting study

But…

(1st, 2nd, 3rd-order)

Model	4th-order ToM	Hi-ToM
GPT-4o	46.17%	69.00%
GPT-4o-mini	30.50%	58.50%
DeepSeek-v3	**58.67%**	**70.17%**
Qwen2.5-0.5B-Instruct	23.83%	30.33%
Qwen2.5-1.5B-Instruct	25.17%	40.67%
Qwen2.5-3B-Instruct	27.50%	39.17%
Qwen2.5-7B-Instruct	28.83%	52.17%
Qwen2.5-7B-Instruct-1M	17.83%	40.67%
Qwen2.5-0.5B-Instruct (RL)	85.83%	70.83%
Qwen2.5-1.5B-Instruct (RL)	89.33%	79.17%
Qwen2.5-3B-Instruct (RL)	88.17%	81.17%
Qwen2.5-7B-Instruct (RL)	82.83%	83.33%
Qwen2.5-7B-Instruct-1M (RL)	**94.50%**	**84.50%**
Qwen2.5-0.5B-Instruct (SFT)	88.17%	81.00%
Qwen2.5-1.5B-Instruct (SFT)	86.17%	80.50%
Qwen2.5-3B-Instruct (SFT)	92.67%	87.00%
Qwen2.5-7B-Instruct (SFT)	**94.00%**	**87.33%**
Qwen2.5-7B-Instruct-1M (SFT)	93.67%	86.50%

[Lu et al. arXiv'25]

What We Observed in ToM?

- **Our key takeaway**:
  Why do all ToM **post-trained models** perform better on 4th-order ToM than on lower orders, while all **non–post-trained models** do not?

- Higher-orc

1st Where doe
2nd Where doe
3rd Where does Alex think Sally thinks Anne thinks the milk is?

We reproduced, same results!

Model	4th-order ToM	(1st, 2nd, 3rd-order) Hi-ToM
GPT-4o	46.17%	69.00%
GPT-4o-mini	30.50%	58.50%
DeepSeek-v3	**58.67%**	**70.17%**
Qwen2.5-0.5B-Instruct	23.83%	30.33%
Qwen2.5-1.5B-Instruct	25.17%	40.67%
Qwen2.5-3B-Instruct	27.50%	39.17%
Qwen2.5-7B-Instruct	28.83%	52.17%
	17.83%	40.67%
	85.83%	70.83%
	89.33%	79.17%
	88.17%	81.17%
	82.83%	83.33%
	94.50%	**84.50%**
	88.17%	81.00%
	86.17%	80.50%
Qwen2.5-3B-Instruct (SFT)	92.67%	87.00%
Qwen2.5-7B-Instruct (SFT)	**94.00%**	**87.33%**
Qwen2.5-7B-Instruct-1M (SFT)	93.67%	86.50%

[He et al. EMNLP-Findings'23] [Lu et al. arXiv'25]

We Found Shortcuts in ToM Benchmarks

- In the Hi-ToM dataset, a **shortcut**

Shortcuts lead to a false expectation about model capabilities, which is a serious safety issue!

1st	Where does Anne think the milk is?
2nd	Where does Sally think Anne thinks the milk is?
3rd	Where does Alex think Sally thinks Anne thinks the milk is?

Model	4th-order ToM	Hi-ToM (1st, 2nd, 3rd-order)
Qwen2.5-7B-Instruct (RL)	82.83%	83.33%
Qwen2.5-7B-Instruct-1M (RL)	**94.50%**	**84.50%**
Qwen2.5-0.5B-Instruct (SFT)	88.17%	81.00%
Qwen2.5-1.5B-Instruct (SFT)	86.17%	80.50%
Qwen2.5-3B-Instruct (SFT)	92.67%	87.00%
Qwen2.5-7B-Instruct (SFT)	**94.00%**	**87.33%**
Qwen2.5-7B-Instruct-1M (SFT)	93.67%	86.50%

[He et al. EMNLP-Findings'23] [Lu et al. arXiv'25]

We Found Shortcuts in ToM Benchmarks

- We conduct **the first systematic examination** of shortcuts for existing ToM datasets

- **(1) LLM-guided rules**: Simply asking an advanced LLM to discover potential shortcuts, which works well

- **(2) Lexical associations**: Check spurious lexical associations

We Found Shortcuts in ToM Benchmarks

- We audit 8 widely used ToM datasets with different question types
- Narrative vs. Conversational
- State Tracking vs. Intention
- Language Only vs. Vision & Language

Examples of ToM Datasets

- Hi-ToM: narrative, tracking, language only

HI-TOM One-Chapter Story	
1	Emma, Charlotte, Benjamin, Aiden and Isabella entered the workshop.
2	The pear is in the red_treasure_chest.
3	Emma moved the pear to the blue_suitcase.
4	Emma exited the workshop.
5	Charlotte exited the workshop.
6	Benjamin lost his watch.
7	Benjamin exited the workshop.
8	Aiden moved the pear to the blue_crate.
9	Aiden exited the workshop.
10	Isabella moved the pear to the red_treasure_chest.
11	Isabella likes the red_box.
12	Isabella exited the workshop.
13	Aiden publicly claimed that the pear is in the blue_drawer now.
14	Emma privately told Isabella that the radish is in the red_suitcase now.

Order	Question
0th	Where is *O* really?
1st	Where does *A1* think *O* is?
2nd	Where does *A2* think *A1* thinks *O* is?
3rd	Where does *A3* think *A2* thinks *A1* thinks *O* is?
4th	Where does *A4* think *A3* thinks *A2* thinks *A1* thinks *O* is?

[He et al. EMNLP-Findings'23]

Examples of ToM Datasets

- ToMATO: conversational, intention, language only

Conversation
Liam Johnson: “Ethan, I’ve noticed you’ve been a bit distant lately, is everything okay?” Ethan Johnson: “Yeah, everything’s fine, just been really busy with the restaurant and some personal stuff, you know how it is.” Liam Johnson: “Come on, Ethan, we’ve been friends for years, I can tell when something’s bothering you. What’s really going on?” Ethan Johnson: “Alright, alright... to be honest, I’ve been having some trouble making ends meet, the restaurant’s been slow and I’m falling behind on some bills, but I’ll figure it out, I always do.” Liam Johnson: “Ethan, you’re not going to ’figure it out’ on your own, you need a plan and some guidance. Let me take a look at your finances and we can go over some options together, okay?” Ethan Johnson: “Man, I appreciate that, Liam, but I don’t want to be a burden, I can handle it, I just need to get my head out of my... okay, fine, yeah, maybe a fresh set of eyes would help, thanks.” ...
Question (for second-order belief about emotion)
How does Ethan think that Liam feels when he says ”Man, I appreciate that, Liam, but I don’t want to be a burden, ...”?
Options
A: He thinks that he feels concerned and genuinely wants to help, but also might be a bit worried about getting involved in his problems B: He thinks that he feels a sense of determination and seriousness, like he’s taking charge of the situation and wants him to focus on getting back on track C: He thinks that he feels a mix of concern and annoyance, like he’s seen this coming and is a bit exasperated that he didn’t come to him sooner D: He thinks that he feels a sense of warmth and friendship, like he’s happy to be able to help him out and is trying to make him feel better about the situation
Answer: C

[Shinoda et al. AAAI’25]

Examples of ToM Datasets

- MMToM: narrative, both tracking & intention, vision & language

[Jin et al. ACL'24]

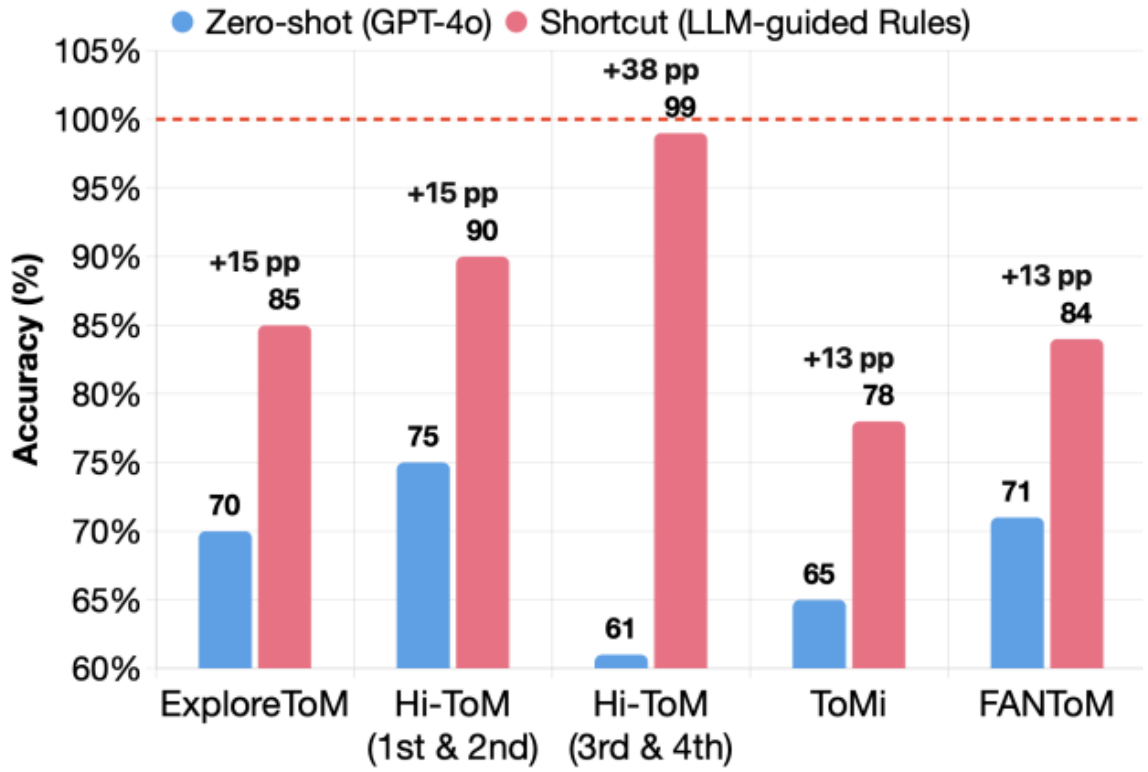
We Found Shortcuts in ToM Benchmarks

- We provide a comprehensive investigation of 8 ToM datasets
- We found that **state tracking problems** are highly shortcut-prone, while **intention problems** require real ToM reasoning

ToM Datasets	Format	Vision	Tracking	Intention	SC (Causal)	SC (Lexical)
ExploreToM	narrative		✓		✓	
FANToM	conversational		✓			✓
ToMi	narrative		✓		✓	✓
Hi-ToM	narrative		✓		✓	✓
OpenToM	narrative		✓	✓		
ToMATO	conversational			✓		
MMToM	narrative	✓	✓	✓		
MuMA-ToM	narrative	✓	✓	✓		

We Found Shortcuts in ToM Benchmarks

- The shortcut issue is serious in 4 out of 8 audited benchmarks
- The **shortcut solution** largely outperforms the **state-of-the-art LLM**



Shortcut Learning Gives a False Sense of ToM

- All the datasets used in [Lu et al. 2025] are shortcut-prone datasets

- → Their findings (e.g., SFT >= RL) may not be true

Model	4th-order ToM	Hi-ToM	ToMi	ExploreToM (Raw)	ExploreToM (Infilled)
GPT-4o	46.17%	69.00%	61.96%	67.35%	62.48%
GPT-4o-mini	30.50%	58.50%	**70.64%**	**69.32%**	**66.04%**
DeepSeek-v3	**58.67%**	**70.17%**	57.17%	65.01%	64.73%
Qwen2.5-0.5B-Instruct	23.83%	30.33%	29.38%	60.51%	54.97%
Qwen2.5-1.5B-Instruct	25.17%	40.67%	54.12%	54.78%	43.53%
Qwen2.5-3B-Instruct	27.50%	39.17%	47.78%	47.37%	49.81%
Qwen2.5-7B-Instruct	28.83%	52.17%	54.65%	59.38%	45.40%
Qwen2.5-7B-Instruct-1M	17.83%	40.67%	54.85%	37.34%	41.46%
Qwen2.5-0.5B-Instruct (RL)	85.83%	70.83%	54.25%	**93.34%**	72.61%
Qwen2.5-1.5B-Instruct (RL)	89.33%	79.17%	75.89%	90.06%	70.73%
Qwen2.5-3B-Instruct (RL)	88.17%	81.17%	80.18%	93.43%	**78.14%**
Qwen2.5-7B-Instruct (RL)	82.83%	83.33%	73.99%	91.65%	74.77%
Qwen2.5-7B-Instruct-1M (RL)	**94.50%**	**84.50%**	**81.08%**	92.31%	77.20%
Qwen2.5-0.5B-Instruct (SFT)	88.17%	81.00%	77.79%	89.68%	69.89%
Qwen2.5-1.5B-Instruct (SFT)	86.17%	80.50%	76.33%	93.53%	74.67%
Qwen2.5-3B-Instruct (SFT)	92.67%	87.00%	79.55%	95.78%	74.95%
Qwen2.5-7B-Instruct (SFT)	**94.00%**	**87.33%**	80.85%	**95.97%**	**77.95%**
Qwen2.5-7B-Instruct-1M (SFT)	93.67%	86.50%	**81.10%**	95.12%	75.61%

[Lu et al. arXiv'25]

Robust Evaluation of ToM Post-Training

- We experiment with the 4 **shortcut-free datasets** that cover different scenarios: OpenToM (narrative), ToMATO (conversational), and MMToM / MuMA-ToM (vision & language)
- **Thinking RFT** > SFT > No-Thinking RFT > Zero-shot

OpenToM (narrative)

Method	Loc (Cg)		Loc (Fg)		MH		Att.	Avg(1^{st}/2^{st}/overall$_{\Delta \text{ vs. SFT}}$)
	First	Second	First	Second	First	Second		
Qwen-2.5-3B Models								
Zero-shot	51.00	50.00	28.00	15.00	53.00	48.00	39.00	44.00 / 37.67 / 40.57$_{\downarrow \mathbf{39.90}}$
SFT	100.00	81.00	92.00	56.00	88.00	76.00	49.00	93.33 / 71.00 / 77.43
No-Thinking RFT	81.00	50.00	81.00	63.00	56.00	55.00	35.00	72.67 / 56.00 / 60.14$_{\downarrow \mathbf{17.29}}$
Thinking RFT	99.00	88.00	94.00	67.00	91.00	85.00	57.00	94.67 / 80.00 / 83.00$_{\uparrow \mathbf{5.57}}$

Robust Evaluation of ToM Post-Training

- RFT enables larger gains on mind-state related questions (e.g., desire and intention)

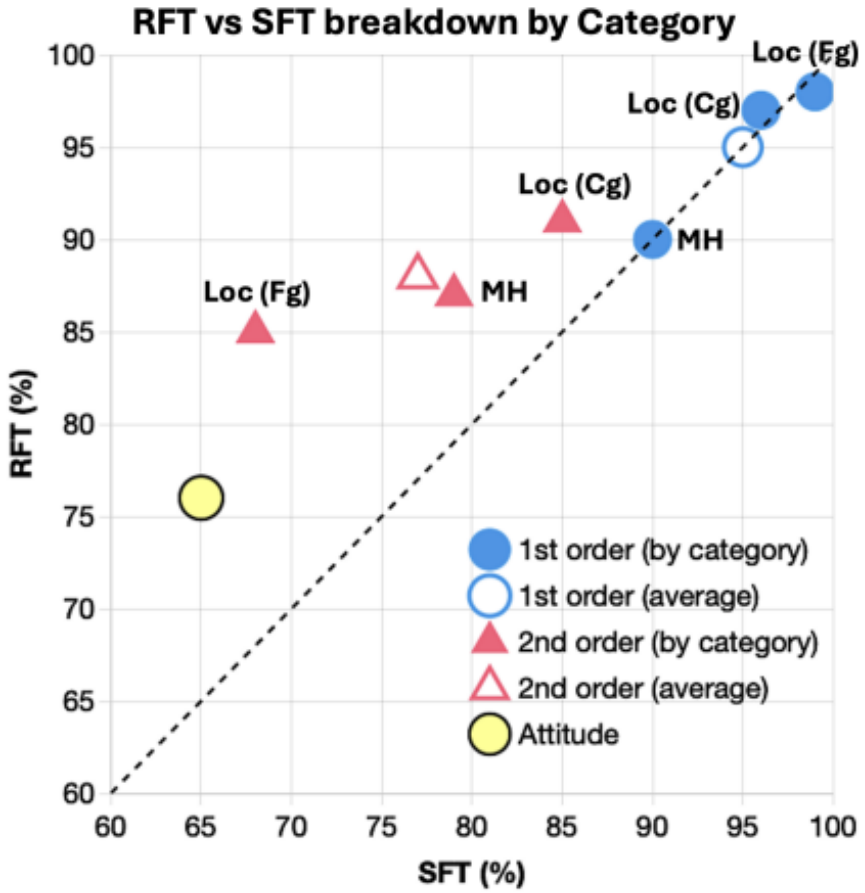
ToMATO (conversational)

Method	Belief	Desire	Emotion	Intention	Knowledge	Avg$_{\Delta \text{ vs. SFT}}$
Zero-shot	65.71	71.63	67.88	65.77	66.51	67.50$_{\downarrow \mathbf{20.42}}$
SFT	87.17	89.07	90.05	86.94	87.05	87.92
No-Thinking RFT	85.18	89.53	84.91	87.61	88.21	87.09$_{\downarrow \mathbf{0.83}}$
Thinking RFT	88.50	92.33	88.81	90.54	89.86	90.00$_{\uparrow \mathbf{2.08}}$

Robust Evaluation of ToM Post-Training

- RFT excels in complex scenarios such as mind-state related questions (attitude) and higher-order reasoning

- Entries above the diagonal line denote where RFT performs better than SFT



Robust Evaluation of ToM Post-Training

- RFT excels in complex scenarios such as multimodal inputs

MMToM & MuMA-ToM (conversational)

Method	Train Modality	MMToM	MuMA-ToM	Avg$_{\Delta \text{ vs. SFT}}$
Zero-shot	Lan.	39.4	–	–
Zero-shot	Lan.+Vis.	45.00	43.30	44.15$_{\downarrow 30.60}$
SFT	Lan.	73.02	–	–
SFT	Lan.+Vis.	74.30	75.20	74.75
Thinking-RFT	Lan.	78.50	–	–
Thinking-RFT	Lan.+Vis.	83.30	81.10	82.20$_{\uparrow 7.45}$

Robust Evaluation of ToM Post-Training

- RFT has better **generalization** from lower- to higher-order ToM compared to SFT

Method	**First Order** (*Seen*)		↪ **Second Order** (*Unseen*)	
	OpenToM	ToMATO	OpenToM	ToMATO
Zero-shot	50.67	72.96	45.67	62.22
SFT	93.00	88.08	$65.33_{\downarrow 27.67}$	$81.74_{\downarrow 6.34}$
RFT	**95.00**	**89.32**	$\mathbf{74.33}_{\downarrow \mathbf{20.67}}$	$\mathbf{84.78}_{\downarrow \mathbf{4.54}}$

Robust Evaluation of ToM Post-Training

- RFT has better **cross-dataset generalization**
- Training dataset: OpenToM
- Test dataset: ToMATO & ExploreToM

Method	ToMATO	ExploreToM
Zero-shot	67.5	62.0
SFT	56.8	63.5
Thinking-RFT	70.4	71.0

In Contrast, If We Didn't Find Shortcuts…

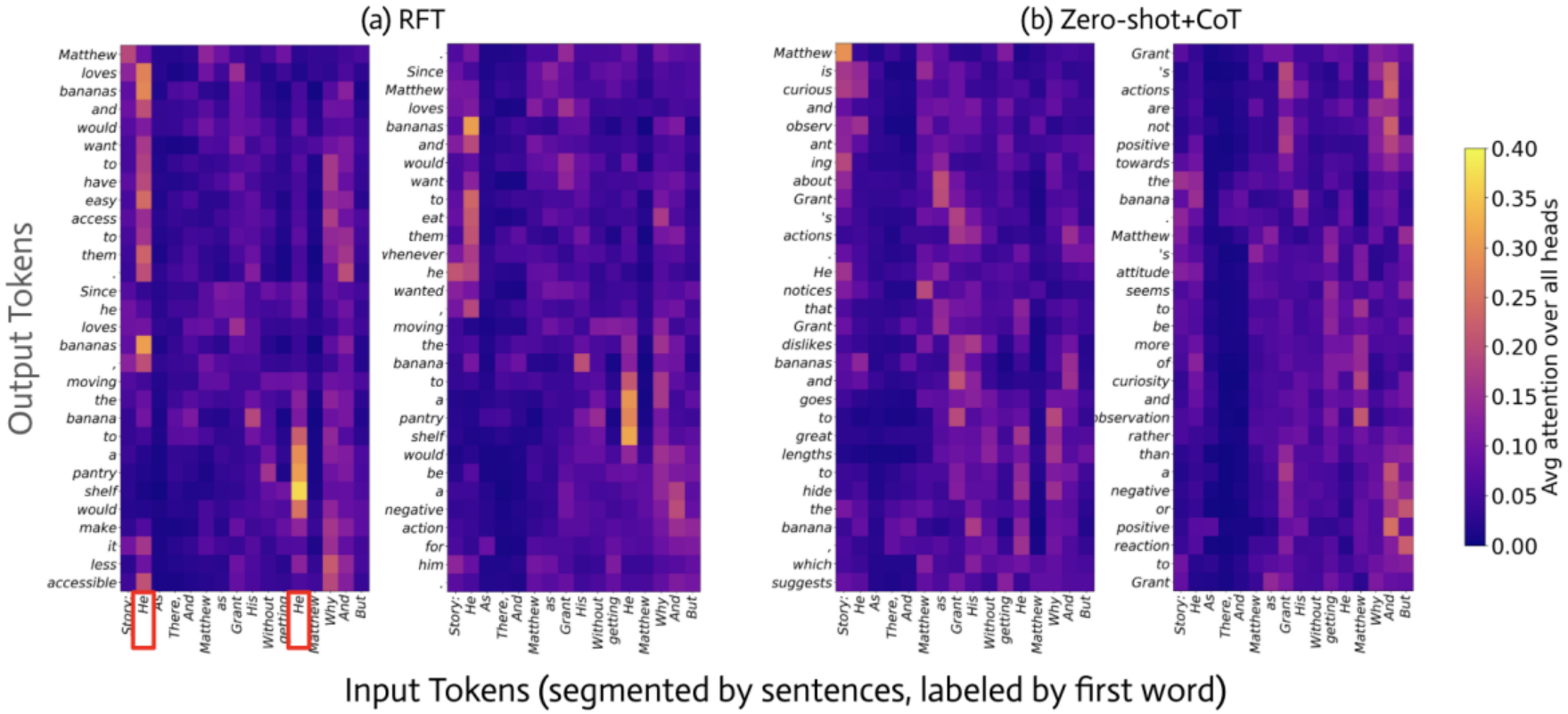
- Train on a shortcut-prone dataset: ExploreToM (in-domain)
- OOD test dataset: Hi-ToM

	3B model		7B model	
Method	In-domain	OOD	In-domain	OOD
Zero-shot	49.5	38.5	62.0	43.6
SFT	96.4	32.5	95.8	34.2
Thinking-RFT	93.2	31.3	94.3	35.3
No-Thinking-RFT	95.8	32.0	96.1	34.0

- Shortcut learning gives a false sense of ToM capabilities
- Shortcut learning inverts the effectiveness of training methods
- Shortcut learning masks the benefits of model scaling
- Shortcut learning harms model generalization

Attention Visualization

- **RFT** model aligns more closely with key information
- In contrast, the **zero-shot+CoT** model exhibits unfocused attention



Thinking-RFT Reasoning Traces

- Quantify the quality of reasoning traces via LLM-as-a-judge
- LC: Logical Consistency; F: Faithfulness; E: Efficiency (max score = 10)

- Feeding only the Thinking-RFT reasoning traces to a frozen base model (zero-shot) can significantly increase accuracy

Method	OpenToM		ToMATO	
	Acc	LLM-Judge (LC/F/E)	Acc	LLM-Judge (LC/F/E)
Zero-shot	46.4	4.3 / 2.2 / 8.0	67.5	5.6 / 4.2 / 7.6
Thinking-RFT	89.1	9.1 / 9.9 / 6.5	90.0	9.2 / 10.0 / 7.0
Zero-shot + RFT Reasoning Trace	74.7	–	82.0	–

Summary

- We find that ToM explicitly benefits from reasoning-based RL: **Thinking RFT** > SFT > No-Thinking RFT > Zero-shot

- Our findings help prevent future ToM research from heading in the wrong direction

- We hope these findings serve as guidelines for designing future ToM benchmarks

Lesson

- Exam your datasets carefully
- Avoid shortcut learning and the false expectations it creates
- Ensure robust model capabilities and safe deployment