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

3/27/2026

Outline

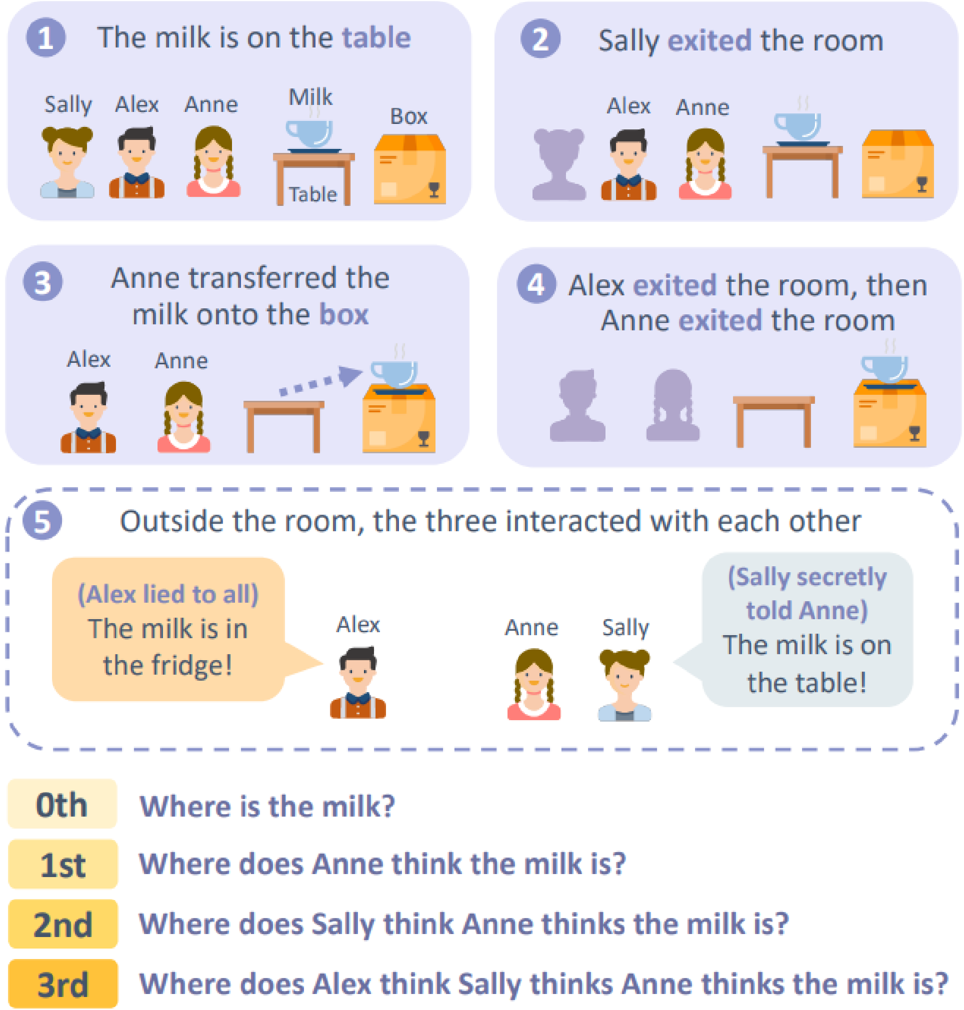
- What is Theory of Mind (ToM)?
- What is shortcut learning?
- Shortcut learning in ToM
- Robust evaluation of ToM post-training

- 年度代表字！

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

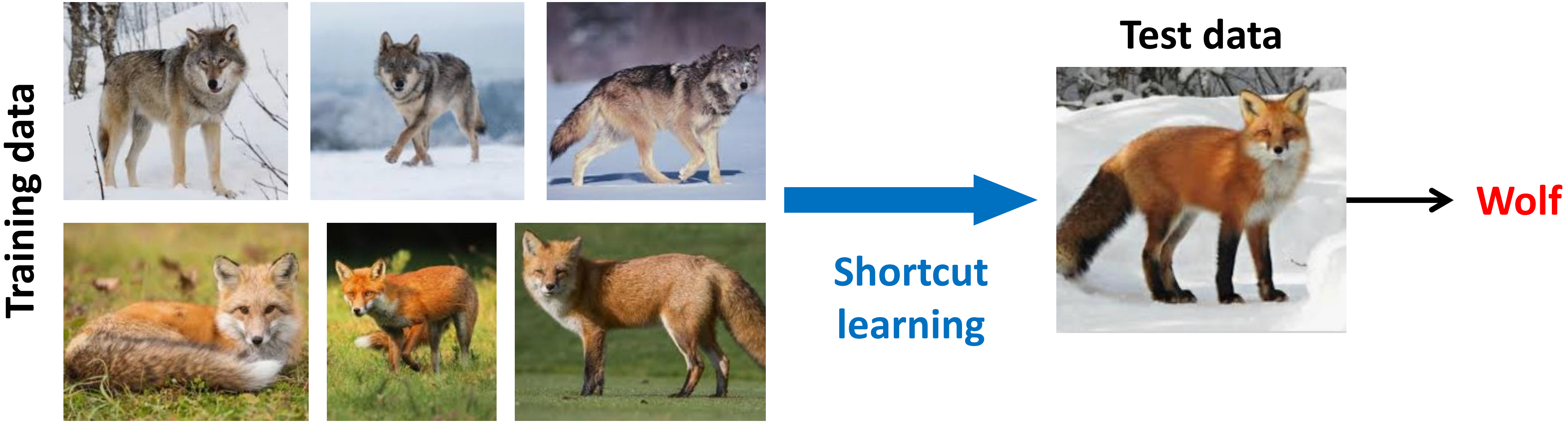
- Infer user’s mental states, such as thoughts, intentions, and beliefs
- Track perspectives (what the human knows vs. doesn’t know)
- Predict next actions and anticipate needs

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

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

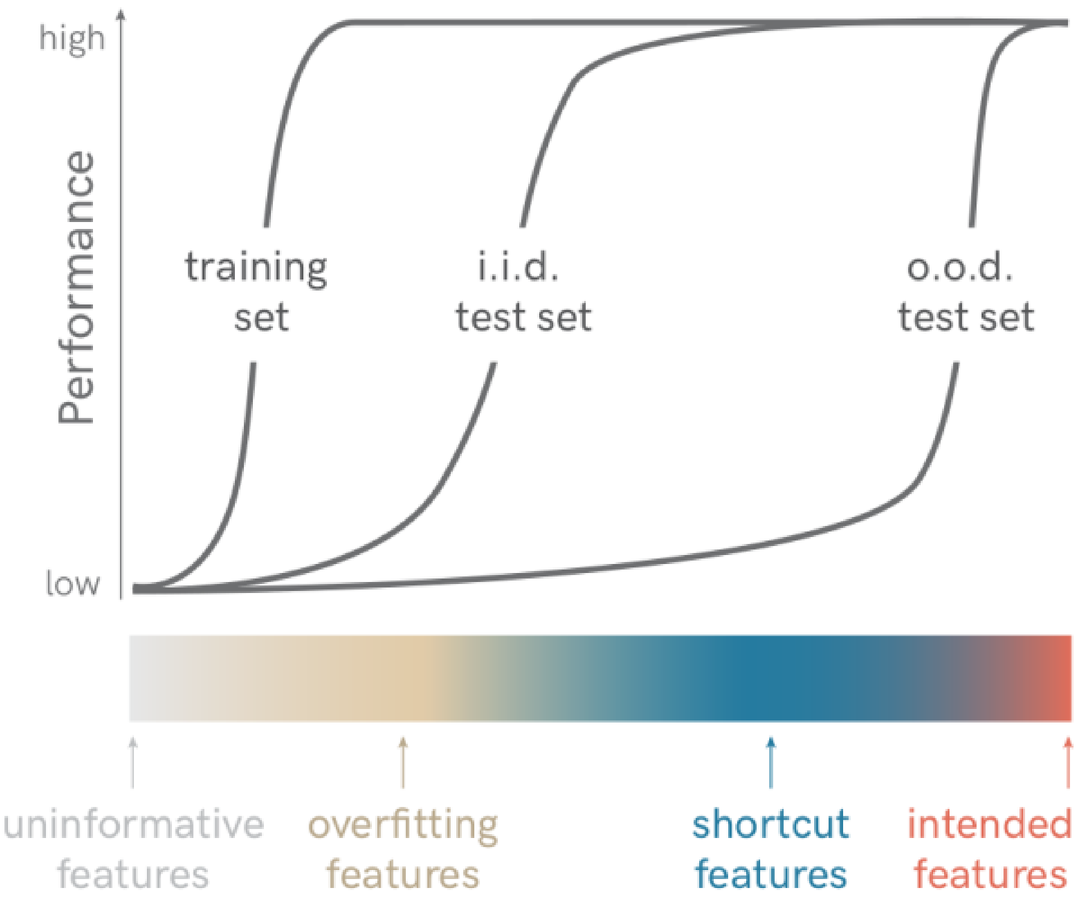
What is Shortcut Learning?

- 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-ord

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	Hi-ToM (1st, 2nd, 3rd-order)
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**

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%

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?

[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

VIDEO INPUT

TEXT INPUT

What's inside the apartment: … The kitchen is equipped with a microwave, eight cabinets, … Inside the microwave, there is a cupcake. There is a wine glass and an apple on one of the kitchen tables. There are water glasses, a bottle wine, a condiment bottle, and a bag of chips in inside the cabinets. …
Actions taken by Emily: Emily is initially in the bathroom. She then walks to the kitchen, goes to the sixth cabinet, opens it, subsequently closes it, and then goes towards the fourth cabinet.

QUESTION

Which one of the following statements is more likely to be true?

(a) Emily has been trying to get a cupcake. ✔ (b) Emily has been trying to get a wine glass. ✖

[Jin et al. ACL'24]

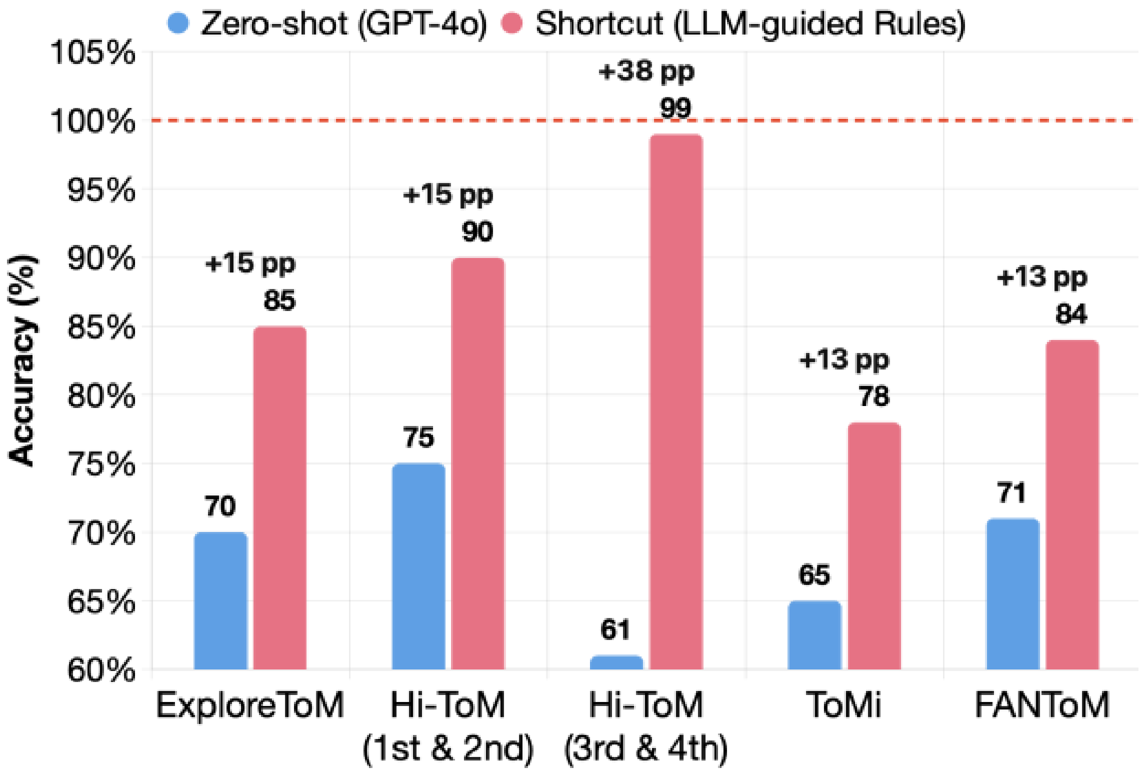
We Found Shortcuts in ToM Benchmarks

- 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$_{\downarrow 39.90}$
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$_{\downarrow 17.29}$
Thinking RFT	99.00	88.00	94.00	67.00	91.00	85.00	57.00	94.67 / 80.00 / 83.00$_{\uparrow 5.57}$

Robust Evaluation of ToM Post-Training

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

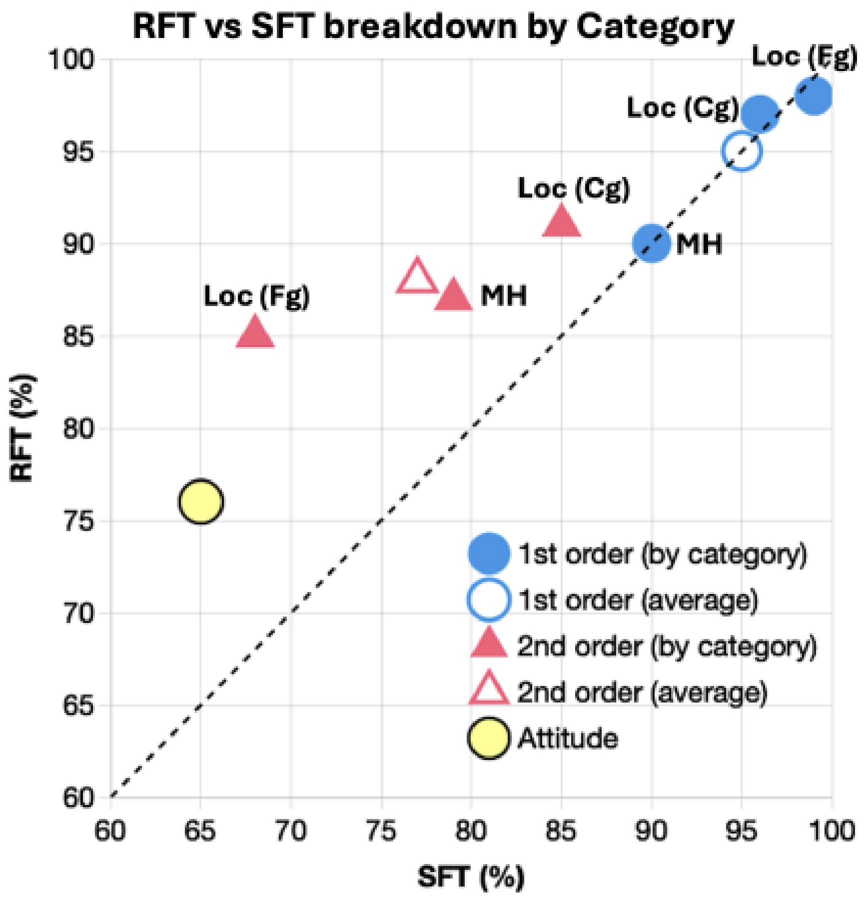
ToMATO (conversational)

Method	Belief	Desire	Emotion	Intention	Knowledge	$\text{Avg}_{\Delta \text{ vs. SFT}}$
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	$\text{Avg}_{\Delta \text{ vs. SFT}}$
Zero-shot	Lan.	39.4	–	–
Zero-shot	Lan.+Vis.	45.00	43.30	$44.15_{\downarrow \mathbf{30.60}}$
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	$82.20_{\uparrow \mathbf{7.45}}$

Robust Evaluation of ToM Post-Training

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

Method	**First Order** (*Seen*)		$\hookrightarrow$ **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**	$\mathbf{74.33}_{\downarrow 20.67}$	$\mathbf{84.78}_{\downarrow 4.54}$

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…

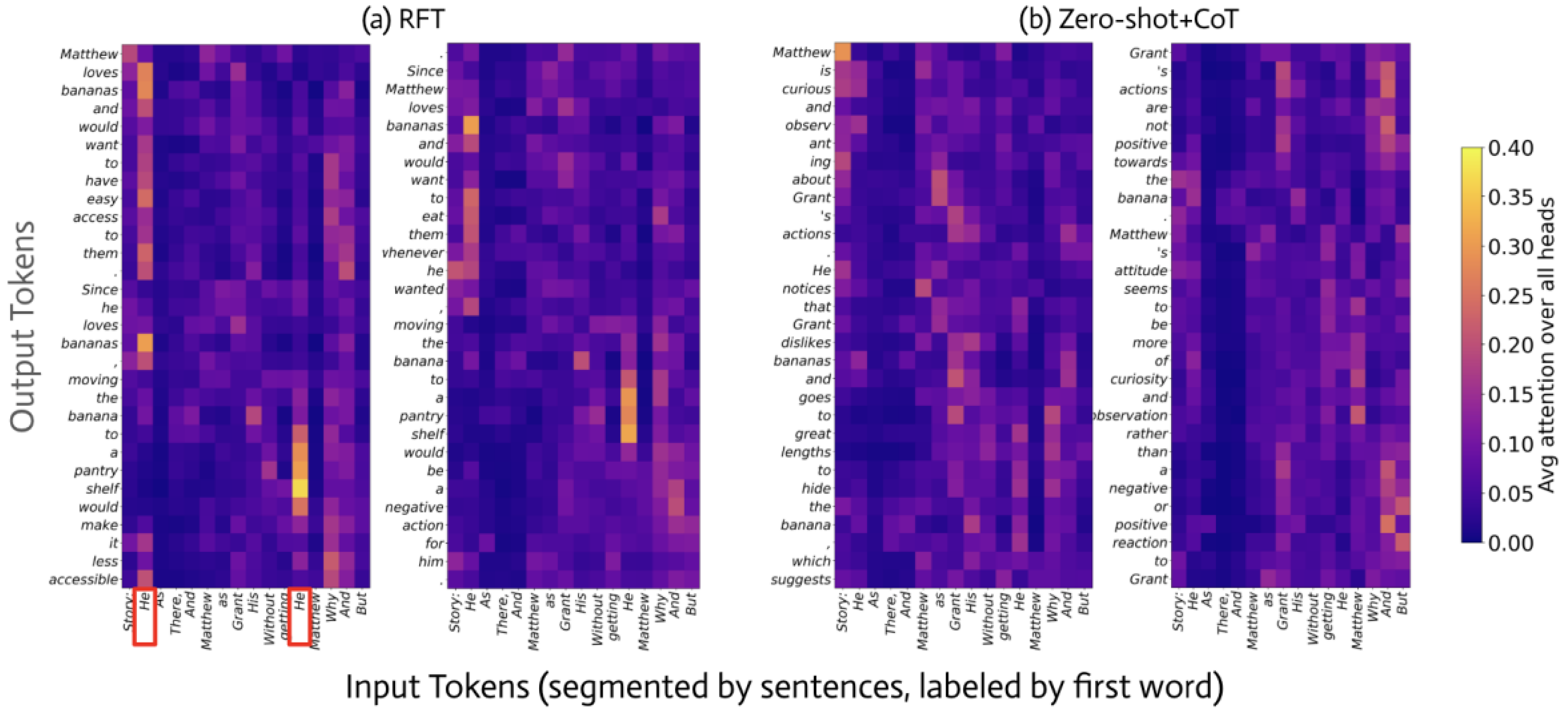
- Train on a shortcut-prone dataset: ExploreToM (in-domain)
- OOD test dataset: Hi-ToM

	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

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

USC

Honda Research Institute US

Jike Zhong

Behzad Dariush

年度代表字！

2016	苦
2017	茫
2018	翻
2019	亂
2020	疫

2021	宅
2022	漲
2023	缺
2024	貪
2025	罷

[Source: Wikipedia]

年度代表字！CV/ML

- 2020 Self-Supervised Learning (MoCo/SimCLR)
- 2021 Vision Transformers (ViT)
- 2022 Vision-Language Pre-Training (CLIP)
- 11/2022 ChatGPT
- 2023 Vision Foundation Models (SAM)
- 2024 Multimodal Large Language Models (LLaVA)
- 2025 Reasoning Models (DeepSeek-R1)
- 2026 ?

2020: Self-Supervised Learning

[Source: www.youtube.com/@DrOsbert]

2020: Self-Supervised Learning

11/2019: **MoCo** (CVPR’20 Best Paper Nominee) 19,499 citations

Momentum Contrast for Unsupervised Visual Representation Learning

Kaiming He Haoqi Fan Yuxin Wu Saining Xie Ross Girshick

Facebook AI Research (FAIR)

Code: https://github.com/facebookresearch/moco

02/2020: **SimCLR** (ICML’20) 30,782 citations

A Simple Framework for Contrastive Learning of Visual Representations

Ting Chen [1] Simon Kornblith [1] Mohammad Norouzi [1] Geoffrey Hinton [1]

2021: Vision Transformers

10/2020: **ViT** (ICLR'21) 90,315 citations

AN IMAGE IS WORTH 16X16 WORDS:
TRANSFORMERS FOR IMAGE RECOGNITION AT SCALE

Alexey Dosovitskiy[*,†], **Lucas Beyer**[*], **Alexander Kolesnikov**[*], **Dirk Weissenborn**[*],
Xiaohua Zhai[*], **Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer,**
Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, Neil Houlsby[*,†]
[*]equal technical contribution, [†]equal advising
Google Research, Brain Team
{adosovitskiy, neilhoulsby}@google.com

2021: Vision Transformers

Medical Transformer: Gated Axial-Attention for Medical Image Segmentation

Jeya Maria Jose Valanarasu[1], Poojan Oza[1], Ilker Hacihaliloglu[2], and Vishal M. Patel[1]

[1] Johns Hopkins University, Baltimore, MD, USA
[2] Rutgers, The State University of New Jersey, NJ, USA

SegFormer: Simple and Efficient Design for Semantic Segmentation with Transformers

Enze Xie[1] **Wenhai Wang**[2] **Zhiding Yu**[3*] **Anima Anandkumar**[3,4] **Jose M. Alvarez**[3] **Ping Luo**[1]
[1]The University of Hong Kong [2]Nanjing University [3]NVIDIA [4]Caltech

xieenze@hku.hk, wangwenhai362@163.com,
{zhidingy, josea, aanandkumar}@nvidia.com, pluo@cs.hku.hk

TransWeather: Transformer-based Restoration of Images Degraded by Adverse Weather Conditions

Jeya Maria Jose Valanarasu, Rajeev Yasarla, and Vishal M. Patel
Johns Hopkins University
{jvalana1,ryasarl1,vpatel36} @jhu.edu

HyperTransformer: A Textural and Spectral Feature Fusion Transformer for Pansharpening

Wele Gedara Chaminda Bandara, Vishal M. Patel
Johns Hopkins University
Department of Electrical and Computer Engineering, Baltimore, MD 21218, USA
{wbandar1, vpatel36}@jhu.edu

A TRANSFORMER-BASED SIAMESE NETWORK FOR CHANGE DETECTION

Wele Gedara Chaminda Bandara, Vishal M. Patel

Johns Hopkins University, Baltimore, Maryland, USA.
{wbandar1, vpatel36}@jhu.edu

2022: Vision-Language Pre-Training

02/2021: **CLIP** (ICML'21) 57,195 citations

Learning Transferable Visual Models From Natural Language Supervision

---

Alec Radford [*1] Jong Wook Kim [*1] Chris Hallacy [1] Aditya Ramesh [1] Gabriel Goh [1] Sandhini Agarwal [1]
Girish Sastry [1] Amanda Askell [1] Pamela Mishkin [1] Jack Clark [1] Gretchen Krueger [1] Ilya Sutskever [1]

2022: Vision-Language Pre-Training

CLIP goes 3D: Leveraging Prompt Tuning for Language Grounded 3D Recognition

Deepti Hegde*, Jeya Maria Jose Valanarasu*, Vishal M. Patel
Johns Hopkins University
dhegde1,jvalana1,vpatel36@jhu.edu

PointCLIP: Point Cloud Understanding by CLIP

Renrui Zhang[*1,3], Ziyu Guo[*2], Wei Zhang[1], Kunchang Li[1], Xupeng Miao[2]
Bin Cui[2], Yu Qiao[1], Peng Gao[†1], Hongsheng Li[3,4]
[1]Shanghai AI Laboratory
[2]School of CS and Key Lab of HCST, Peking University
[3]CUHK-SenseTime Joint Laboratory, The Chinese University of Hong Kong
[4]Centre for Perceptual and Interactive Intelligence (CPII)
{zhangrenrui, gaopeng}@pjlab.org.cn hsli@ee.cuhk.edu.hk

Learning to Prompt CLIP for Monocular Depth Estimation: Exploring the Limits of Human Language

Dylan Auty, Imperial College London, dylan.auty12@imperial.ac.uk
Krystian Mikolajczyk, Imperial College London, k.mikolajczyk@imperial.ac.uk

ActionCLIP: Adapting Language-Image Pretrained Models for Video Action Recognition

Mengmeng Wang, Jiazheng Xing, Jianbiao Mei, *Graduate Student Member, IEEE*,
Yong Liu, *Member, IEEE*, and Yunliang Jiang

Simple Open-Vocabulary Object Detection with Vision Transformers

Matthias Minderer*, Alexey Gritsenko*,
Austin Stone, Maxim Neumann, Dirk Weissenborn, Alexey Dosovitskiy,
Aravindh Mahendran, Anurag Arnab, Mostafa Dehghani, Zhuoran Shen,
Xiao Wang, Xiaohua Zhai, Thomas Kipf, and Neil Houlsby

Google Research
{mjlm,agritsenko}@google.com

Scaling Open-Vocabulary Image Segmentation with Image-Level Labels

Golnaz Ghiasi, Xiuye Gu, Yin Cui, and Tsung-Yi Lin*

Google Research
{golnazg, xiuyegu, yincui}@google.com tsungyil@nvidia.com

2023: Vision Foundation Models

04/2023: **SAM** (ICCV’23) 19,396 citations

Segment Anything

Alexander Kirillov[1,2,4] Eric Mintun[2] Nikhila Ravi[1,2] Hanzi Mao[2] Chloe Rolland[3] Laura Gustafson[3]
Tete Xiao[3] Spencer Whitehead Alexander C. Berg Wan-Yen Lo Piotr Dollár[4] Ross Girshick[4]
[1]project lead [2]joint first author [3]equal contribution [4]directional lead
Meta AI Research, FAIR

2024: Multimodal Large Language Models

04/2023: **LLaVA** (NeurIPS'23) 13,258 citations

Visual Instruction Tuning

Haotian Liu[1*], **Chunyuan Li**[2*], **Qingyang Wu**[3], **Yong Jae Lee**[1]
[1]University of Wisconsin–Madison [2]Microsoft Research [3]Columbia University
https://llava-vl.github.io

2024: Multimodal Large Language Models

Towards Zero-Shot Anomaly Detection and Reasoning with Multimodal Large Language Models

Jiacong Xu[1*] Shao-Yuan Lo[2] Bardia Safaei[1] Vishal M. Patel[1] Isht Dwivedi[2]
[1]Johns Hopkins University [2]Honda Research Institute USA
{jxu155, bsafaei1, vpatel36}@jhu.edu {shao-yuan_lo, idwivedi}@honda-ri.com

Can't make an Omelette without Breaking some Eggs: Plausible Action Anticipation using Large Video-Language Models

Himangi Mittal[1,2*] Nakul Agarwal[1] Shao-Yuan Lo[1] Kwonjoon Lee[1]
[1]Honda Research Institute USA [2]Carnegie Mellon University
hmittal@andrew.cmu.edu {nakul_agarwal, shao-yuan_lo, kwonjoon_lee}@honda-ri.com

StimuVAR: Spatiotemporal Stimuli-aware Video Affective Reasoning with Multimodal Large Language Models

Yuxiang Guo[1*†], Faizan Siddiqui[2], Yang Zhao[1], Rama Chellappa[1*], Shao-Yuan Lo[2*]
[1]Johns Hopkins University.
[2]Honda Research Institute USA.

Video-LLaMA
An Instruction-tuned Audio-Visual Language Model for Video Understanding

Hang Zhang[1 2] **Xin Li**[1 2*] **Lidong Bing**[1 2]
[1] DAMO Academy, Alibaba Group
[2] Hupan Lab, 310023, Hangzhou, China
{zh401075, xinting.lx, l.bing}@alibaba-inc.com

LLaVA-Med: Training a Large Language-and-Vision Assistant for Biomedicine in One Day

Chunyuan Li[*], **Cliff Wong**[*], **Sheng Zhang**[*], **Naoto Usuyama, Haotian Liu, Jianwei Yang Tristan Naumann, Hoifung Poon, Jianfeng Gao**
Microsoft
https://aka.ms/llava-med

2025: Reasoning Models

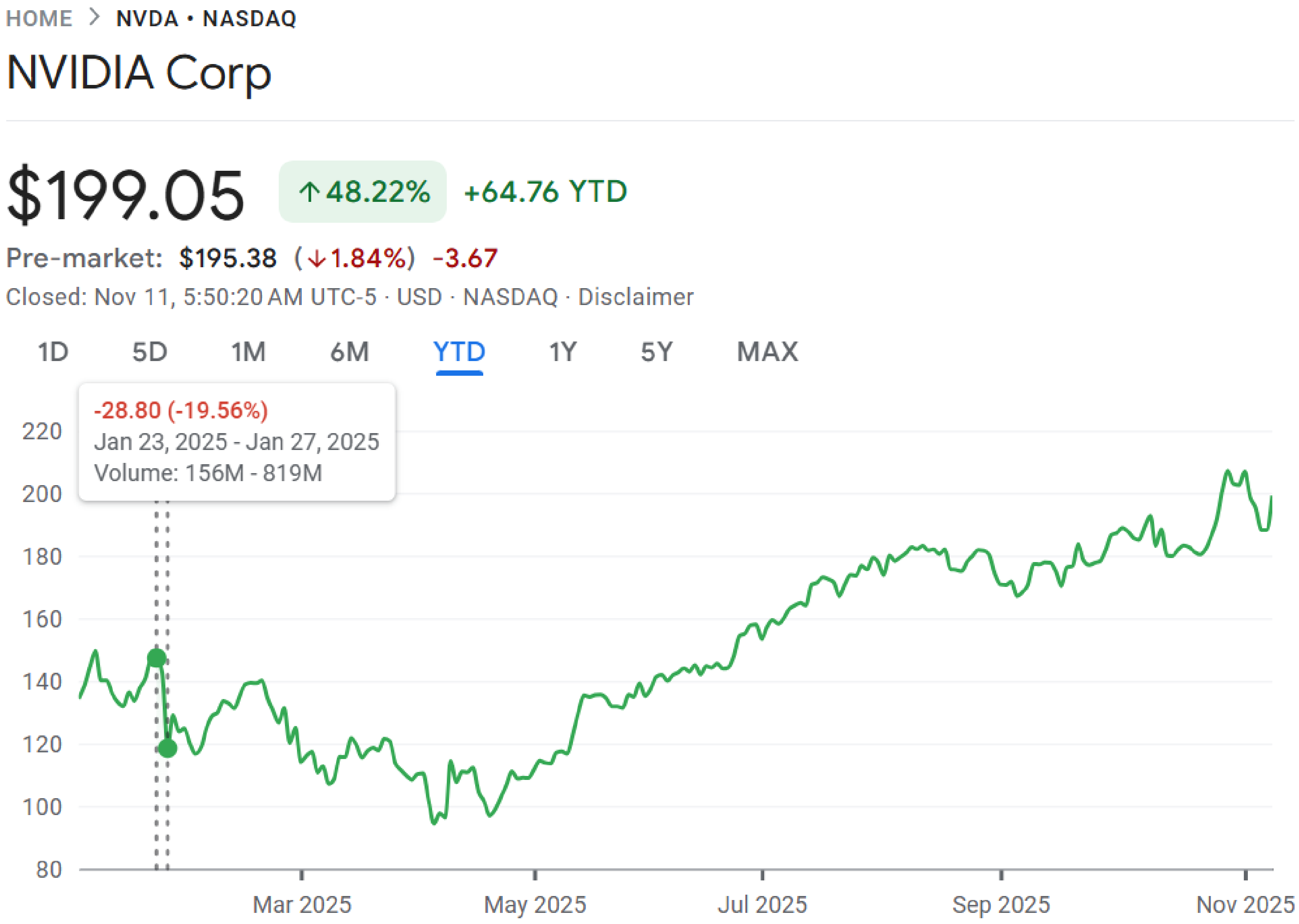
01/2025: **DeepSeek-R1** (Nature’25) 8,417 citations

DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning

DeepSeek-AI

research@deepseek.com

2025: Reasoning Models



2025: Reasoning Models

VISION-R1: INCENTIVIZING REASONING CAPABILITY IN MULTIMODAL LARGE LANGUAGE MODELS

Wenxuan Huang[1,2*†] Bohan Jia[1*] Zijie Zhai[1] Shaosheng Cao[3✉]
Zheyu Ye[3] Fei Zhao[3] Zhe Xu[3] Xu Tang[3] Yao Hu[3] Shaohui Lin[1✉]
[1]East China Normal University [2]The Chinese University of Hong Kong [3]Xiaohongshu Inc.
wxhuang0616@gmail.com (Wenxuan Huang)
*: Equal Contribution †: Project Leader ✉: Corresponding Author

Med-R1: Reinforcement Learning for Generalizable Medical Reasoning in Vision-Language Models

Yuxiang Lai, Jike Zhong, Ming Li, Shitian Zhao, Yuheng Li, Konstantinos Psounis, *Fellow, IEEE*, and Xiaofeng Yang, *Member, IEEE*

Video-R1: Reinforcing Video Reasoning in MLLMs

Kaituo Feng[1], Kaixiong Gong[1], Bohao Li[2], Zonghao Guo[3*], Yibing Wang[4], Tianshuo Peng[1], Junfei Wu[4], Xiaoying Zhang[5], Benyou Wang[2], Xiangyu Yue[1*]
[1]CUHK MMLab, [2]CUHK (SZ), [3]Tsinghua University, [4]UCAS, [5]CUHK HCCL
https://github.com/tulerfeng/Video-R1

IAD-R1: Reinforcing Consistent Reasoning in Industrial Anomaly Detection

Yanhui Li[1], Yunkang Cao[2], Chengliang Liu[3], Yuan Xiong[1], Xinghui Dong[4], Chao Huang[1*]
[1] Sun Yat-sen University [2] Hunan University
[3] University of Macau [4] Ocean University of China
liyh665@mail2.sysu.edu.cn, caoyunkang0207@gmail.com, liucl1996@163.com
{xiongy89, huangch253}@mail.sysu.edu.cn, xinghui.dong@ouc.edu.cn

GEOVLM-R1: REINFORCEMENT FINE-TUNING FOR IMPROVED REMOTE SENSING REASONING

Mustansar Fiaz[1], Hiyam Debary[1], Paolo Fraccaro[1], Danda Paudel[2], Luc Van Gool[2,3], Fahad Khan[4,5], Salman Khan[4,6]
[1]IBM Research, [2]INSAIT, [3]ETH Zürich, [4]MBZUAI, [5]Linköping University, [6]ANU Australia

Vad-R1: Towards Video Anomaly Reasoning via Perception-to-Cognition Chain-of-Thought

Chao Huang[1] Benfeng Wang[1] Wei Wang[1*] Jie Wen[2] Chengliang Liu[3] Li Shen[1,4] Xiaochun Cao[1]
[1]Shenzhen Campus of Sun Yat-sen University [2]Harbin Institute of Technology, Shenzhen
[3]Laboratory for Artificial Intelligence in Design, The Hong Kong Polytechnic University
[4]Shenzhen Loop Area Institute
{huangch253, wangwei29, caoxiaochun}@mail.sysu.edu.cn wangbf23@mail2.sysu.edu.cn
wenjie@hit.edu.cn liucl1996@163.com mathshenli@gmail.com

年度代表字！CV/ML

- 2020 Self-Supervised Learning (MoCo/SimCLR)
- 2021 Vision Transformers (ViT)
- 2022 Vision-Language Pre-Training (CLIP)
- 11/2022 ChatGPT
- 2023 Vision Foundation Models (SAM)
- 2024 Multimodal Large Language Models (LLaVA)
- 2025 Reasoning Models (DeepSeek-R1)
- 2026 Latent Reasoning (Coconut)

2026: Latent Reasoning

12/2024: **Coconut** (COLM'25) 421 citations

Training Large Language Models to Reason in a Continuous Latent Space

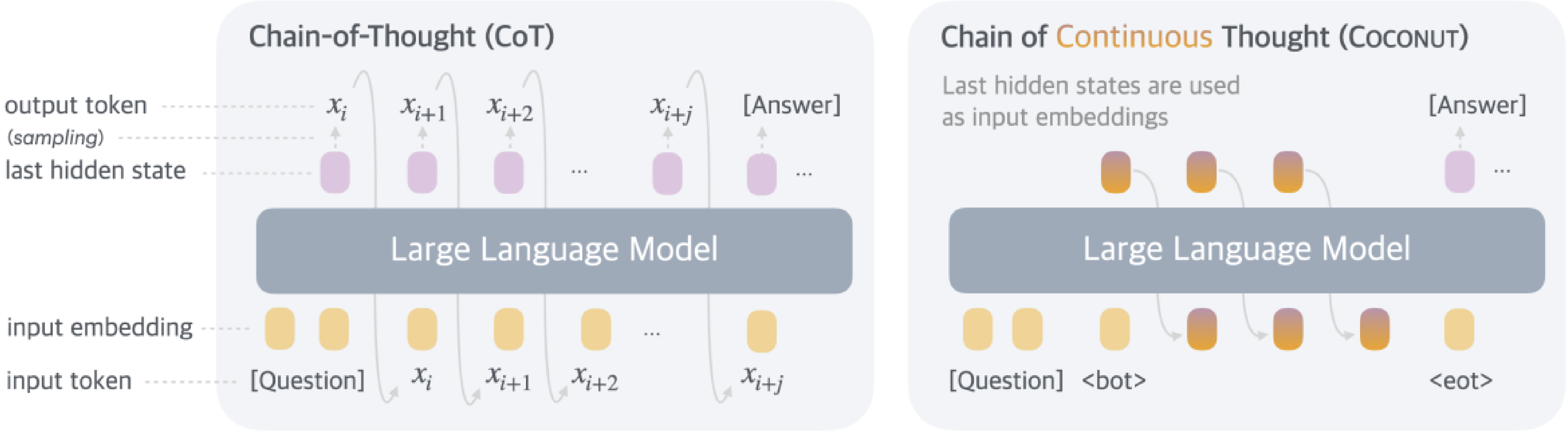
Shibo Hao[1,2,*], **Sainbayar Sukhbaatar**[1], **DiJia Su**[1], **Xian Li**[1], **Zhiting Hu**[2], **Jason Weston**[1], **Yuandong Tian**[1]

[1]FAIR at Meta, [2]UC San Diego

[*]Work done at Meta

2026: Latent Reasoning

- Reasoning in the latent space



[Hao et al. COLM'25]

2026: Latent Reasoning

- Humans do not always think in an **explicit** linguistic format
- Similarly, textual chain-of-thought (CoT) reasoning may not be the most effective internal computation medium for neural networks
- Latent reasoning explores alternative forms that enable more powerful **implicit** reasoning for neural networks

2026: Thinking with Images

04/2025: OpenAI Blog

April 16, 2025 Release

Thinking with images

Try on ChatGPT ↗

2026: Thinking with Images