Sparsity and Scaling: Unveiling and Enhancing Theory-of-Mind in Large Language Models

Yuheng Wu
yuhengwu@stanford.edu
Dec 17 @ ASAP Seminar

Background: Theory-of-Mind (ToM)

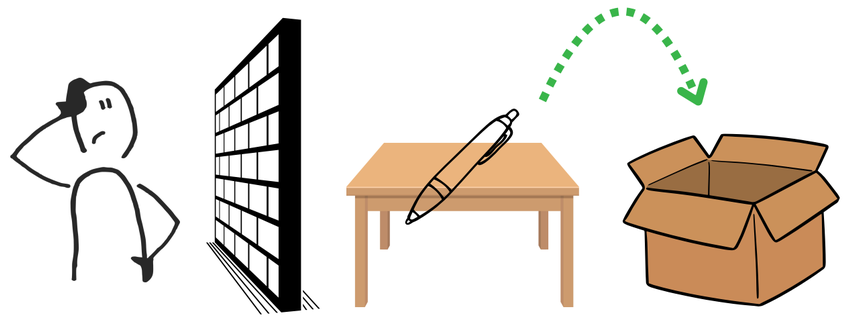
- Theory-of-mind is the cognitive ability to attribute mental states (such as beliefs and intentions) to oneself and others, and to understand that **others have beliefs that are different from one's own**.
- If you can answer the question below correctly…

Theory-of-mind task example (unexpeced transfer)

***Story*:**

1. Alice and Bob are in the living room. A pen is on the table.
2. Alice left the living room. Bob moved the pen to the box.

Question*: *Where does Alice think the pen is?



Background: Theory-of-Mind (ToM)

- Theory-of-mind is the cognitive ability to attribute mental states (such as beliefs and intentions) to oneself and others, and to understand that **others have beliefs that are different from one's own**.
- If you can answer the question below correctly…

Theory-of-mind task example (unexpeced transfer)

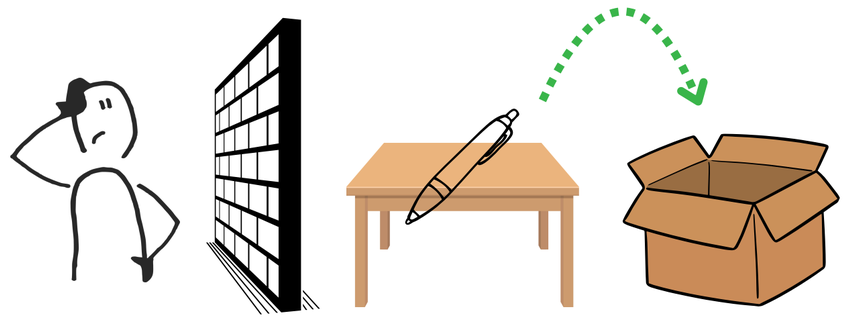
***Story*:**

1. Alice and Bob are in the living room. A pen is on the table.
2. Alice left the living room. Bob moved the pen to the box.

Question*: *Where does Alice think the pen is?

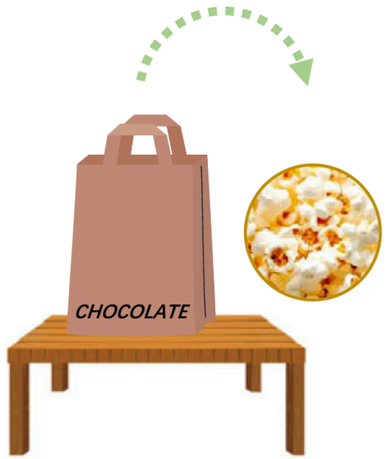
***Conclusion*:**

Alice thinks the pen is on the **table**.



Background: Theory-of-Mind (ToM)

- Significance in Developmental Psychology: Considered a critical milestone in human early cognitive development.
- Classic False Belief Tasks: **Unexpected Transfer** and **Unexpected Contents**.



False Belief ToM Task

① Here is a bag filled with **popcorn**. ② Yet the label on the bag says **chocolate**. ③ Sam finds the bag and reads the label. ④ Sam **doesn't** open the bag and **doesn't** look inside.

Question (a)

Sam **opens** the bag and looks inside. She can clearly see that it is full of ____.

Question (b)

Sam **calls a friend** to tell them that she has just found a bag full of ____.

popcorn

chocolate

child

Michal Kosinski. Evaluating large language models in theory of mind tasks. PNAS (2024).
Strachan et al. Testing theory of mind in large language models and humans. Nature Human Behavior (2024).

Background: Theory-of-Mind (ToM)

- Significance in Developmental Psychology: Considered a critical milestone in human early cognitive development.
- Classic False Belief Tasks: **Unexpected Transfer** and **Unexpected Contents**.
- Advanced Social Scenarios: **Irony** and **faux pas**.

The politician had taken his assistant along to his conference; there were almost no other attendees there.
'Clearly people want to hear you speak', mused the assistant.
Did the assistant think people want to hear the politician speak?

Michal Kosinski. Evaluating large language models in theory of mind tasks. PNAS (2024).
Strachan et al. Testing theory of mind in large language models and humans. Nature Human Behavior (2024).

Background: Theory-of-Mind (ToM)

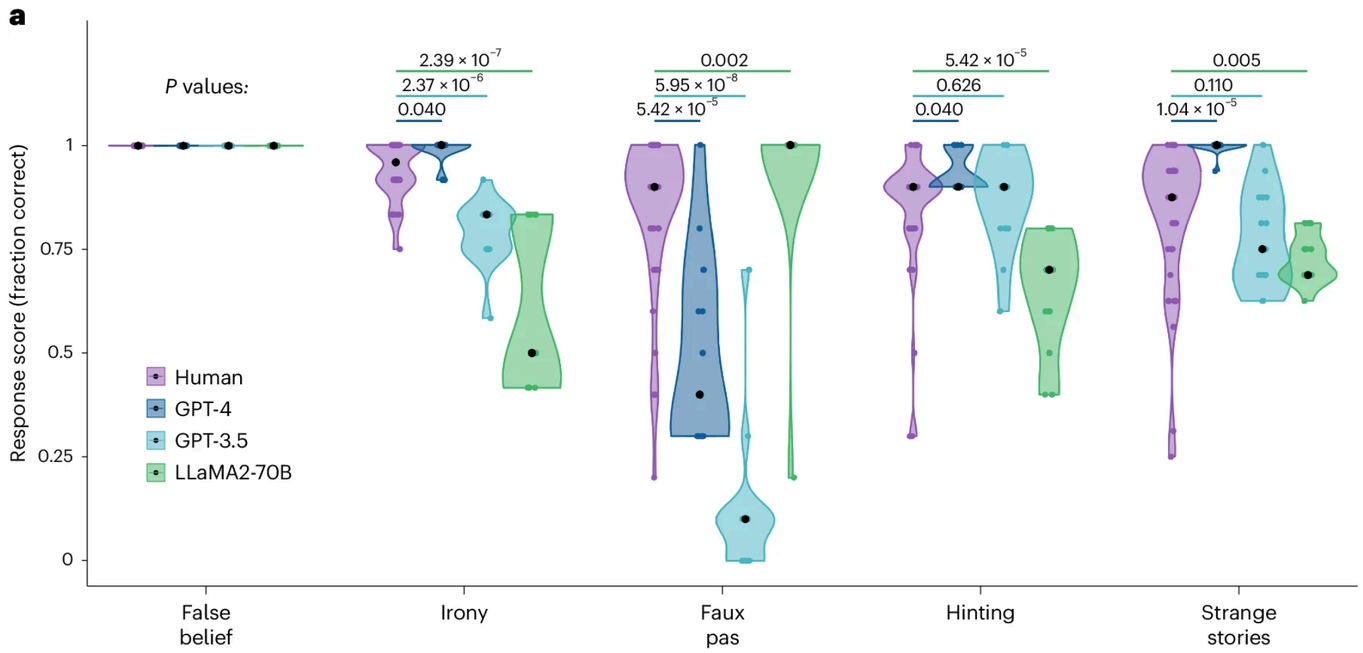
- Significance in Developmental Psychology: Considered a critical milestone in human early cognitive development.
- Classic False Belief Tasks: **Unexpected Transfer** and **Unexpected Contents**.
- Advanced Social Scenarios: **Irony** and **faux pas**.

All of the class took part in a story competition.
Emma really wanted to win. Whilst she was away from school, the results of the competition were announced: Alice was the winner.
The next day, Alice saw Emma and said "I'm sorry about your story." "What do you mean?" said Emma. "Oh nothing," said Alice.
In the story did someone say something that they should not have said?

Michal Kosinski. Evaluating large language models in theory of mind tasks. PNAS (2024).
Strachan et al. Testing theory of mind in large language models and humans. Nature Human Behavior (2024).

Theory-of-Mind (ToM) in LLMs

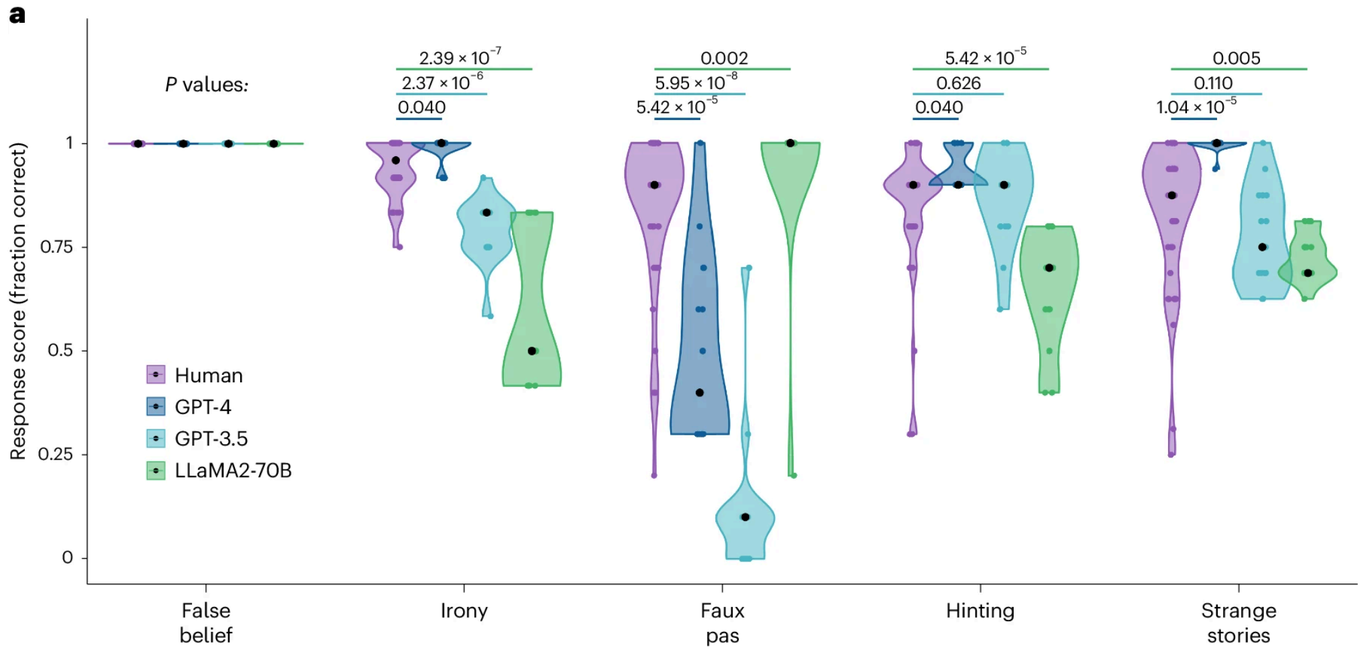
- Do LLMs possess Theory-of-Mind? It seems… **YES**?



Strachan et al. Testing theory of mind in large language models and humans. Nature Human Behavior (2024).

Theory-of-Mind (ToM) in LLMs

- Do LLMs possess Theory-of-Mind? It seems… **YES**?
- Why it matters:ToM is fundamental for building **socially aware agents** capable of **trustworthy and interpretable human-AI interactions**.



Strachan et al. Testing theory of mind in large language models and humans. Nature Human Behavior (2024).

Research Questions

- **RQ1**: How do LLMs encode Theory-of-Mind capabilities?
  - (Focus: Internal Mechanisms & Sparsity)

- **RQ2**: How can we enhance Theory-of-Mind reasoning in LLMs?
  - (Focus: Inference-Time Scaling)

How do LLMs encode ToM capabilities?

- **Step 1**: Identification
  - We introduce a Hessian-based framework to identify and locate the specific parameters sensitive to ToM reasoning.
- **Step 2**: Impact Analysis
  - We investigate how perturbing these parameters affects ToM performance compared to general language capabilities.
- **Step 3**: Mechanism Exploration
  - We trace the underlying cause by analyzing how these parameters interact with core architectural components, specifically Positional Encoding and Attention Mechanisms.

Method: Identifying ToM-Sensitive Params

- **Hessian-based Sensitivity Analysis**: We approximate parameter importance using the Fisher Information Matrix on ToM datasets to identify which weights affect the loss the most.

$$\mathcal{D}_{\text{ToM}-\text{Train}} = \{(x_i, y_i)\}_{i=1}^{n} \qquad \mathcal{L}(\boldsymbol{\theta}; \mathcal{D}_{\text{ToM}-\text{Train}}) = \frac{1}{n}\sum_{i=1}^{n} \ell(\boldsymbol{\theta}; x_i, y_i)$$

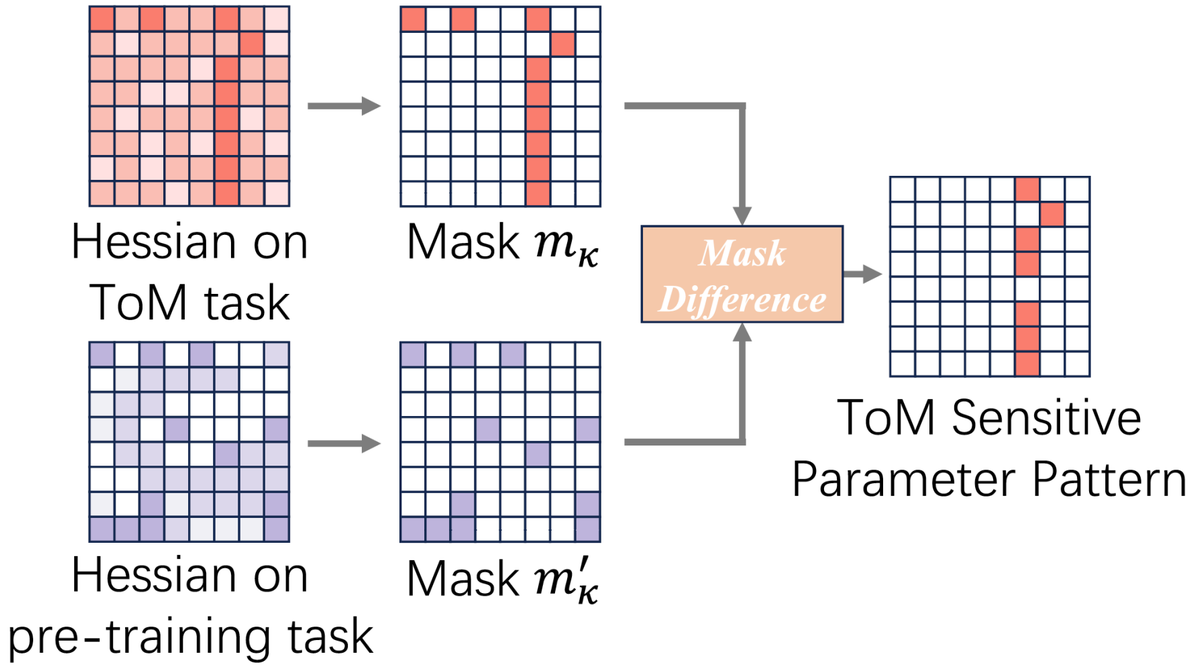
$$\mathbf{g}_i = \nabla_{\boldsymbol{\theta}} \ell(\boldsymbol{\theta};\ x_i,\ y_i)$$

$$\nabla_{\boldsymbol{\theta}} \mathcal{L}(\boldsymbol{\theta}) \approx \frac{1}{n}\sum_{i=1}^{n} \mathbf{g}_i, \quad H \approx F \approx \widehat{F} = \frac{1}{n}\sum_{i=1}^{n} \mathbf{g}_i\, \mathbf{g}_i^{\top}$$

Method: Identifying ToM-Sensitive Params

- **Hessian-based Sensitivity Analysis**: We approximate parameter importance using the Fisher Information Matrix on ToM datasets to identify which weights affect the loss the most.

- **Targeted Isolation**: We compute the intersection to isolate parameters exclusive to ToM reasoning.
- **Goal**: Find parameters where modifying them breaks ToM but keeps the language fluent.



Kim et al. SqueezeLLM: Dense-and-Sparse Quantization. International Conference on Machine Learning (2024).

Result I: Impact on ToM Performance

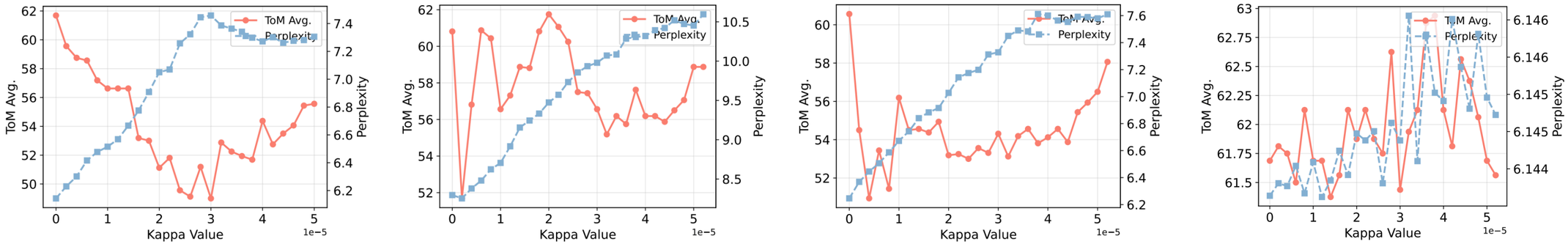
- **Extreme Sparsity, Massive Impact**: Perturbing only **0.001% level** parameters causes a significant drop in ToM task accuracy.
- **Comparison with Random Control**: Randomly perturbing the same number of parameters results in zero degradation.

	Model	Unexpected Contents				Unexpected Transfer				Avg(↑)	PPL(↓)
		FB	CL	IP	OC	FB	NT	IP	PP		
Llama	3-8B	66.00	83.50	94.50	42.00	48.00	63.00	73.00	23.50	61.69	6.14
	3-8B-P	32.00	82.50	81.50	50.00	20.00	50.50	50.50	25.00	49.00	7.46
	3-8B-Ins	87.50	74.00	89.50	41.00	68.00	60.50	47.00	19.00	60.81	8.30
	3-8B-Ins-P	96.00	63.50	66.50	17.00	64.00	60.50	23.00	23.00	51.69	8.25
Qwen	2-7B	50.00	87.50	87.50	75.00	27.50	72.50	75.00	42.50	64.69	7.14
	2-7B-P	52.50	67.50	52.50	40.00	25.00	65.00	50.00	30.00	47.81	7.70
	2-7B-Ins	42.50	85.50	83.50	66.50	24.00	66.00	64.50	38.50	58.88	7.60
	2-7B-Ins-P	47.50	67.00	64.00	38.50	12.00	47.00	43.00	31.50	43.81	8.53

P denotes the version with the sensitive pattern perturbed, and **Ins** represents the Instruct-tuned variant of the model. The abbreviations for ToM tasks are as follows: *FB* False Belief, *CL* Correct Label, *IP* Informed Protagonist, *OC* Open Container, *NT* No Transfer, and *PP* Present Protagonist. Underlined values indicate a decline in model performance after perturbation.

Result I: Impact on ToM Performance

- **Extreme Sparsity, Massive Impact**: Perturbing only **0.001% level** parameters causes a significant drop in ToM task accuracy.
- **Comparison with Random Control**: Randomly perturbing the same number of parameters results in zero degradation.



(a) Llama3-8B (b) Llama3-8B-Instruct (c) Llama3.1-8B (o) Llama3-8B-Random

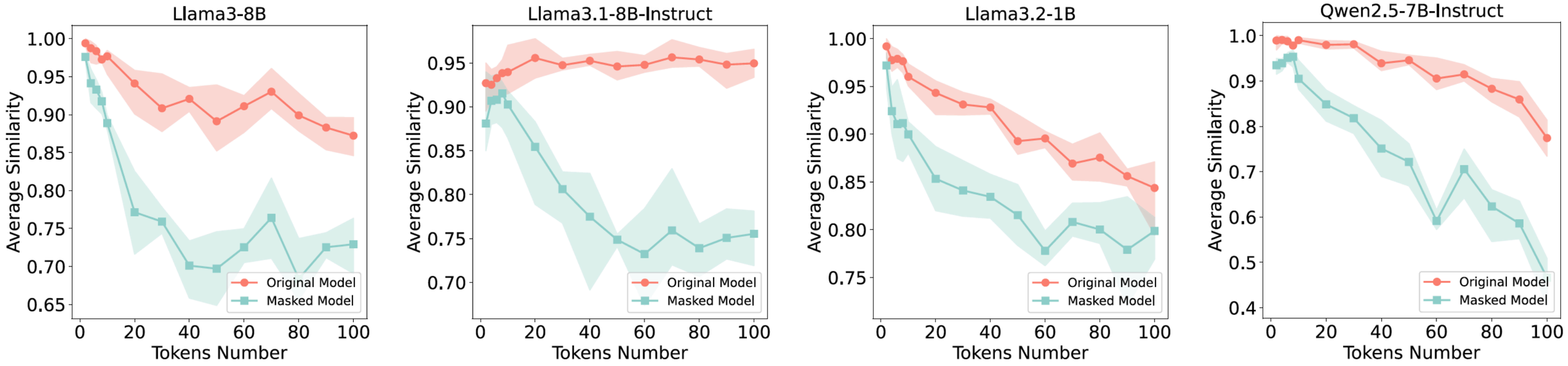
Result I: Impact on ToM Performance

- **Extreme Sparsity, Massive Impact**: Perturbing only **0.001% level** parameters causes a significant drop in ToM task accuracy.
- **Comparison with Random Control**: Randomly perturbing the same number of parameters results in zero degradation.
- **Architecture Dependency**: This sensitivity pattern is consistently observed in RoPE-based models (Llama, Qwen, DeepSeek) but absent in non-RoPE models (e.g., Jamba)

	Model	Unexpected Contents				Unexpected Transfer				Avg(↑)	PPL(↓)
		FB	CL	IP	OC	FB	NT	IP	PP		
Jamba	1.5-Mini	74.00	45.50	93.00	50.50	60.50	65.50	77.50	28.00	61.81	7.77
	1.5-Mini-P	73.00	53.00	90.00	41.00	62.50	77.00	78.50	32.50	63.44	7.67

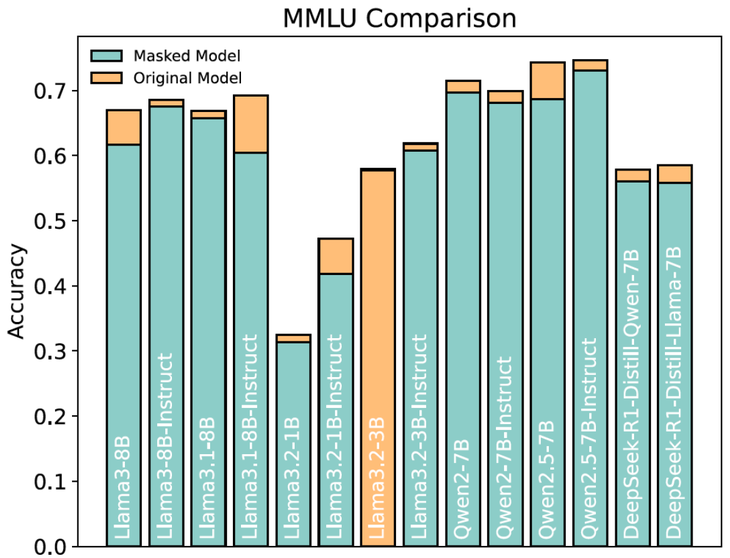
Result II: Impact on Lang. Understanding

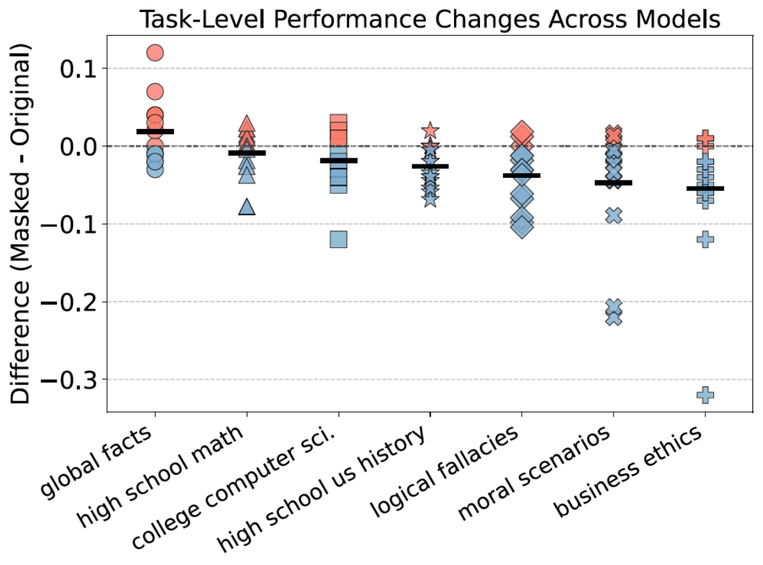
- **Language Fluency Preserved**: PPL remains largely stable. The model retains its ability to generate fluent text.
- **Contextual Localization Fails**: However, the model loses the ability to "locate" information within the context (as seen in repetition/similarity tasks), struggling to retrieve tokens based on position.



Result II: Impact on Lang. Understanding

- **Contextual Localization Fails**: However, the model loses the ability to "locate" information within the context (as seen in repetition/similarity tasks), struggling to retrieve tokens based on position.
- **MMLU Breakdown**: General knowledge declines specifically in subtasks requiring reasoning about human behavior (e.g., Business Ethics).





Findings

- **Findings 1**
  - An extremely **sparse ToM-sensitive parameter pattern** exists, whose perturbation significantly affects **RoPE-based** models' ToM capabilities, while random perturbations do not.
  - Our experiments further demonstrate that this degradation is linked to a reduction in **contextual localization** and **language understanding**.

- Why?

Mechanism I: The RoPE Connection

- **Structural Characteristics**: The identified ToM-sensitive parameters exhibit a low-rank structure.

Table 6: Sensitive parameter mask rank analysis for LLaMA3-8B ($\kappa = 0.000030$)

	$W_{\mathbf{Q}}$	$W_{\mathbf{K}}$	$W_{\mathbf{V}}$	$W_{\mathbf{O}}$	$W_{\mathbf{Gate}}$	$W_{\mathbf{Up}}$	$W_{\mathbf{Down}}$
Original Rank	4020.91	1022.72	1024.00	4083.50	4096.00	4096.00	4096.00
Mask Rank	21.69	10.50	6.88	16.06	29.66	26.09	17.56
Normalized Mask Rank	0.5774	0.6915	0.8512	0.5960	0.3235	0.3002	0.6484

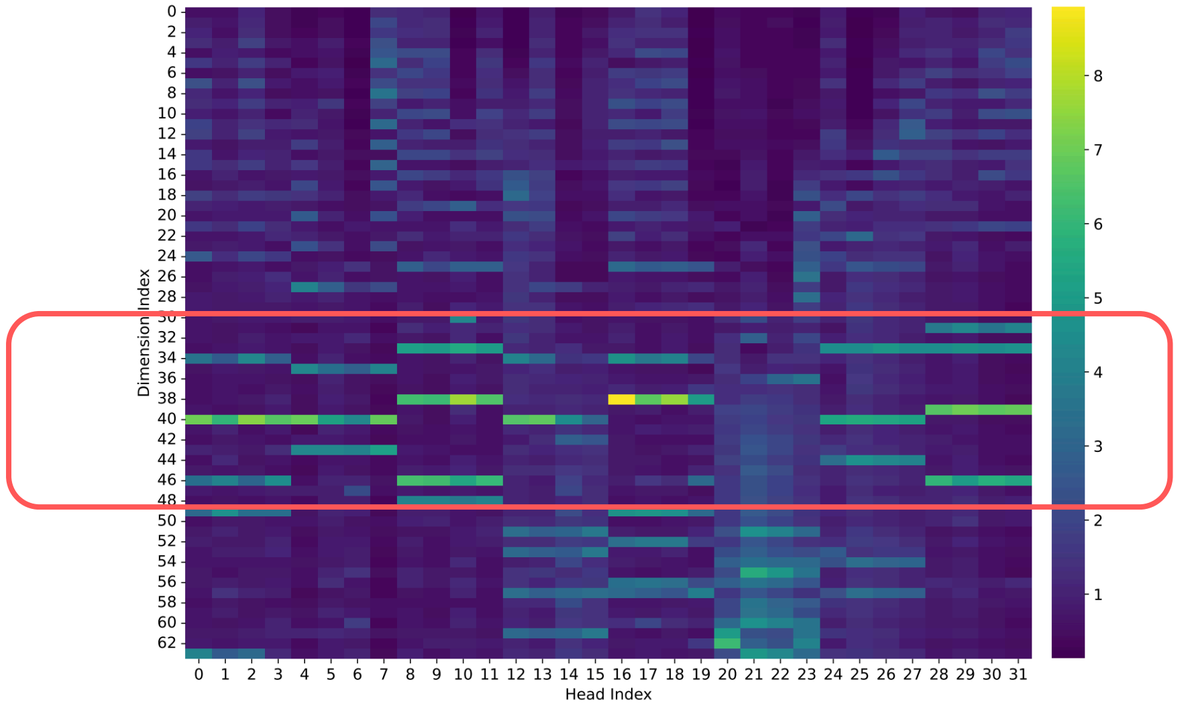
Mechanism I: The RoPE Connection

- **Structural Characteristics**: The identified ToM-sensitive parameters exhibit a low-rank structure.
- **Frequency Alignment**: These parameters align precisely with the dominant frequency activations of RoPE.

$$\theta(p, m) = p \cdot \left(\frac{1}{50000}\right)^{\frac{2m}{d_h}}$$

$$\begin{aligned}\mathrm{Enc}(\mathbf{x}_p^m, p, m) &= \begin{bmatrix} \cos(\theta(p,m)) & -\sin(\theta(p,m)) \\ \sin(\theta(p,m)) & \cos(\theta(p,m)) \end{bmatrix} \cdot \mathbf{x}_p^m \\ &= M(p,m) \cdot \mathbf{x}_p^m.\end{aligned}$$

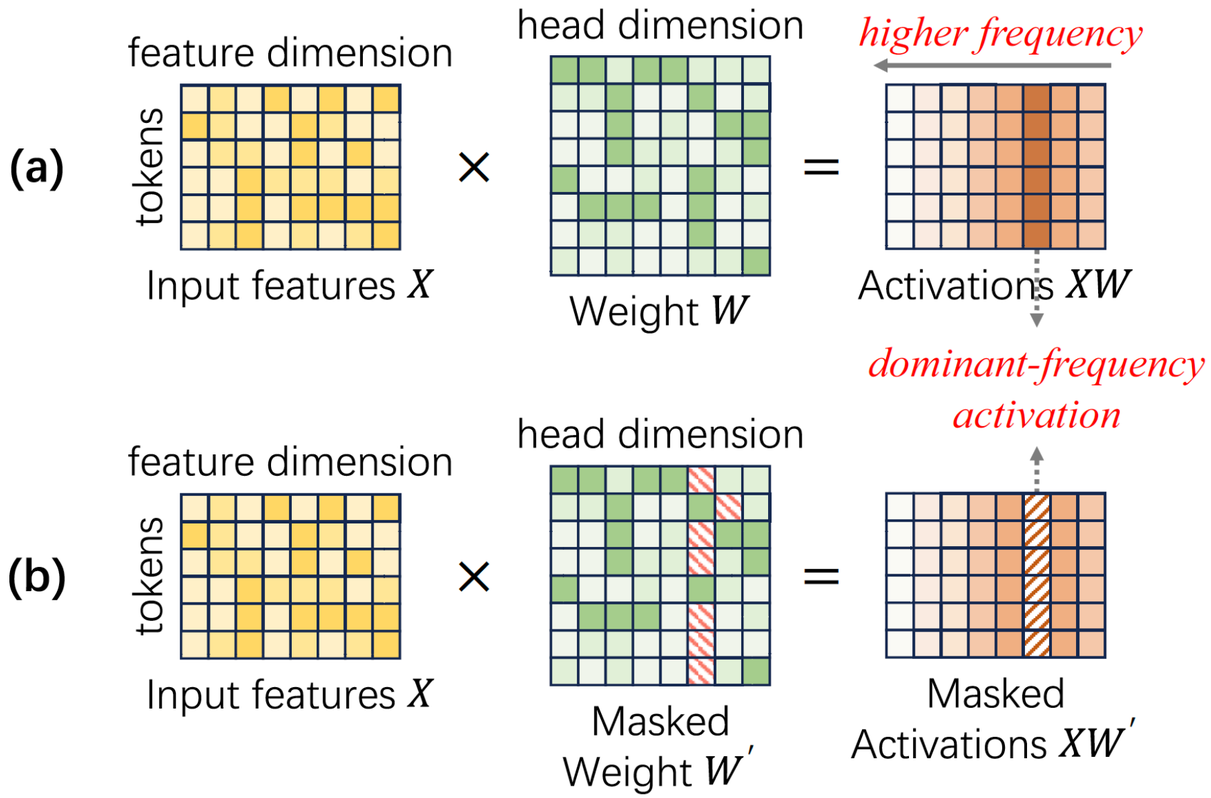
$$\begin{aligned}\mathrm{RoPE}(\mathbf{q}_i, \mathbf{k}_j) &= \sum_{m=0}^{d_h/2-1} \left(\mathrm{Enc}(\mathbf{q}_i^m, i, m)\right)^\top \cdot \mathrm{Enc}(\mathbf{k}_j^m, j, m) \\ &= \sum_{m=0}^{d_h/2-1} (\mathbf{q}_i^m)^\top \cdot M(j-i, m) \cdot \mathbf{k}_j^m.\end{aligned}$$

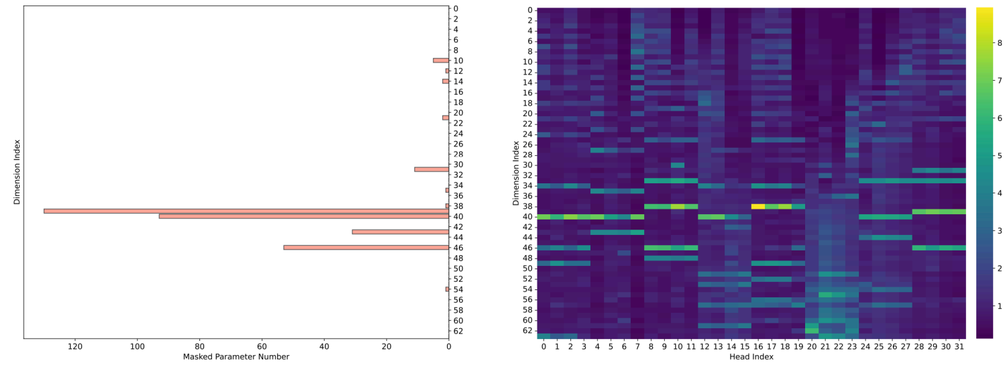


Barbero et al. Round and Round We Go! What makes Rotary Positional Encodings useful? International Conference on Learning Representations (2025).

Mechanism I: The RoPE Connection

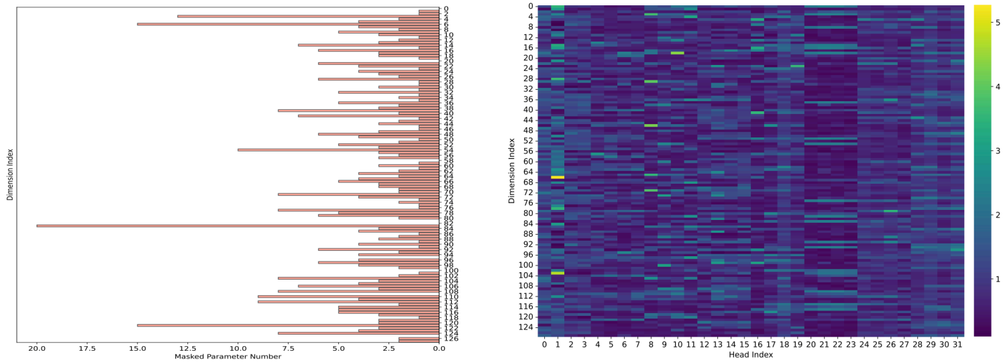
- **The Disruption**: Perturbing these parameters selectively destroys these dominant frequencies.





(a) ToM-sensitive parameter distribution (b) Activation map

Fig. 8: ToM-sensitive parameter distribution and activation map for $W_{\mathbf{Q}}$ matrix in Llama3-8B layer 2.

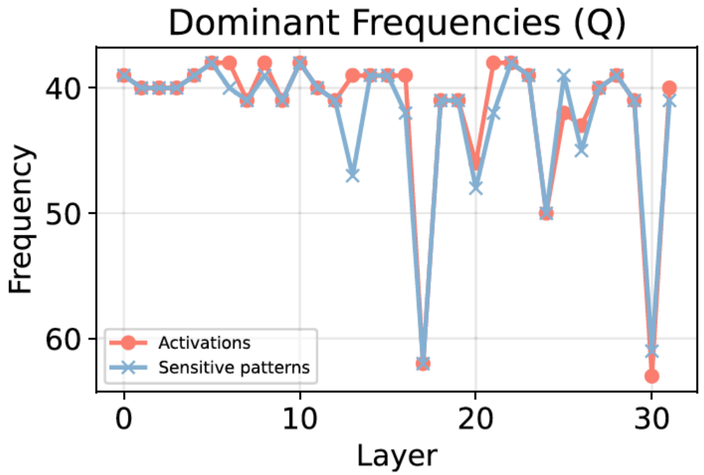


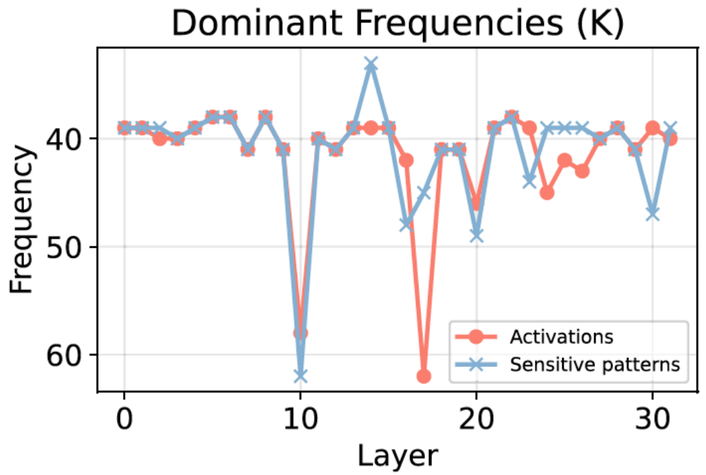
(a) ToM-sensitive parameter distribution (b) Activation map

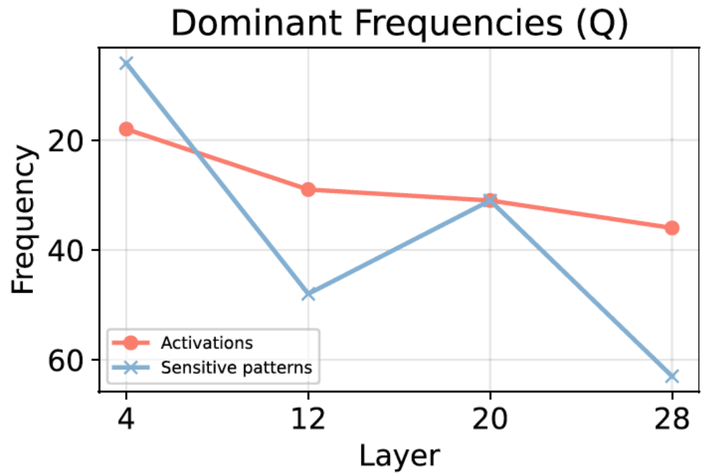
Fig. 9: ToM-sensitive parameter distribution and activation map for $W_{\mathbf{Q}}$ matrix in Jamba-1.5-Mini layer 4.

Mechanism I: The RoPE Connection

- **Structural Characteristics**: The identified ToM-sensitive parameters exhibit a low-rank structure.
- **Frequency Alignment**: These parameters align precisely with the dominant frequency activations of RoPE.
- **The Disruption**: Perturbing these parameters selectively destroys these dominant frequencies.







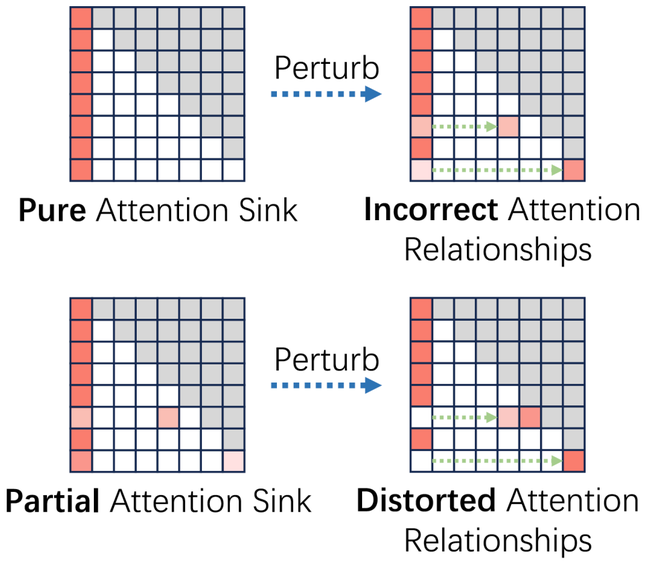
Findings

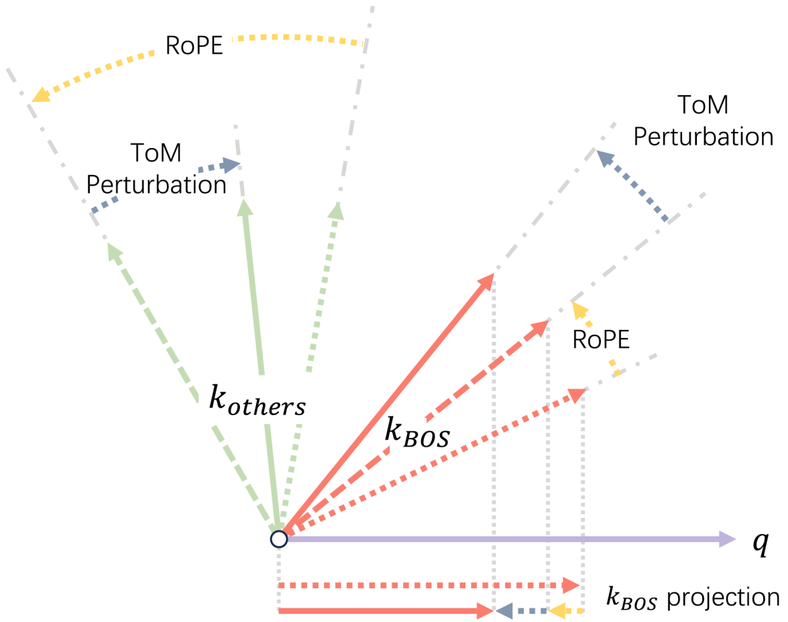
- **Findings 2**
  - The functionality of the ToM-sensitive parameter pattern relates to the **positional encoding** module in LLM architectures.
  - Perturbing the proposed ToM-sensitive parameter pattern in LLMs with **RoPE** disrupts dominant frequency activations induced by positional encoding.
  - In contrast, LLMs without RoPE lack this frequency-dependent activation structure and **exhibit different sensitivity patterns.**

- Then?

Mechanism II: Attention Sink Collapse

- **The Role of Attention Sinks**: Normally, the Beginning-of-Sequence (BOS) token acts as an "Attention Sink," stabilizing the model by absorbing excess attention scores.
- **Geometric Distortion**: Perturbation alters the angle between the Query and the BOS Key, making them nearly orthogonal.



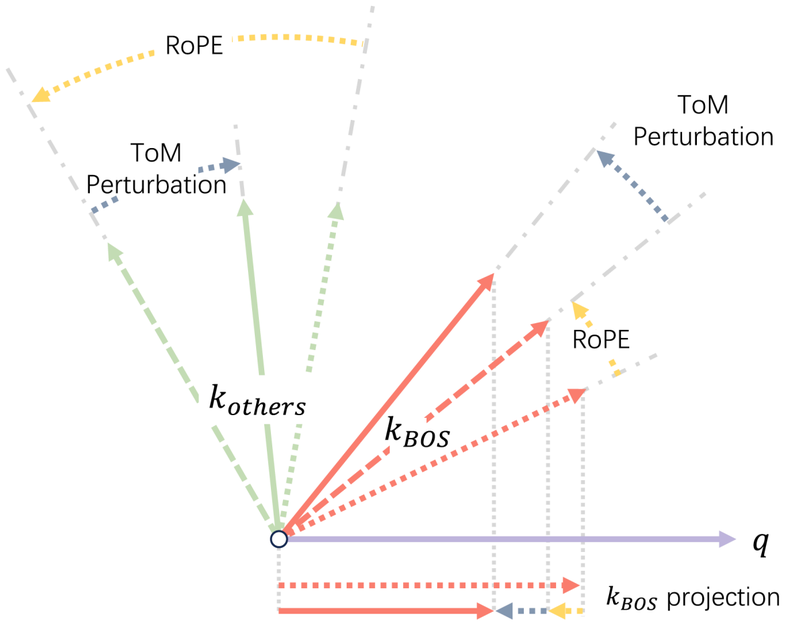


Mechanism II: Attention Sink Collapse

- **Result**: The strong connection to the BOS anchor is severed. Without the sink, attention weights shift significantly to irrelevant tokens.
- **Consequence**: The model loses its ability to maintain stable feature relationships, leading to reasoning failure.

Table 2 | Amplitude and angle of activation embeddings before RoPE, after RoPE, and after Perturbation

	Before RoPE (0)	After RoPE (1)	After Perturb. (2)	Change (0 → 1)	Change (1 → 2)
$\|\mathbf{q}\|_2$	12.95	12.95	12.76	0.00	−0.19
$\|\mathbf{k}_{BOS}\|_2$	4.22	4.22	3.91	0.00	−0.31
$\|\mathbf{k}_{others}\|_2$	22.48	22.48	22.19	0.00	−0.30
$\angle(\mathbf{q}, \mathbf{k}_{BOS})$	66.35	66.46	69.22	0.11	2.77
$\angle(\mathbf{q}, \mathbf{k}_{others})$	93.34	96.81	95.20	3.47	−1.62



Mechanism II: Attention Sink Collapse

- **Result**: The strong connection to the BOS anchor is severed. Without the sink, attention weights shift significantly to irrelevant tokens.
- **Consequence**: The model loses its ability to maintain stable feature relationships, leading to reasoning failure.

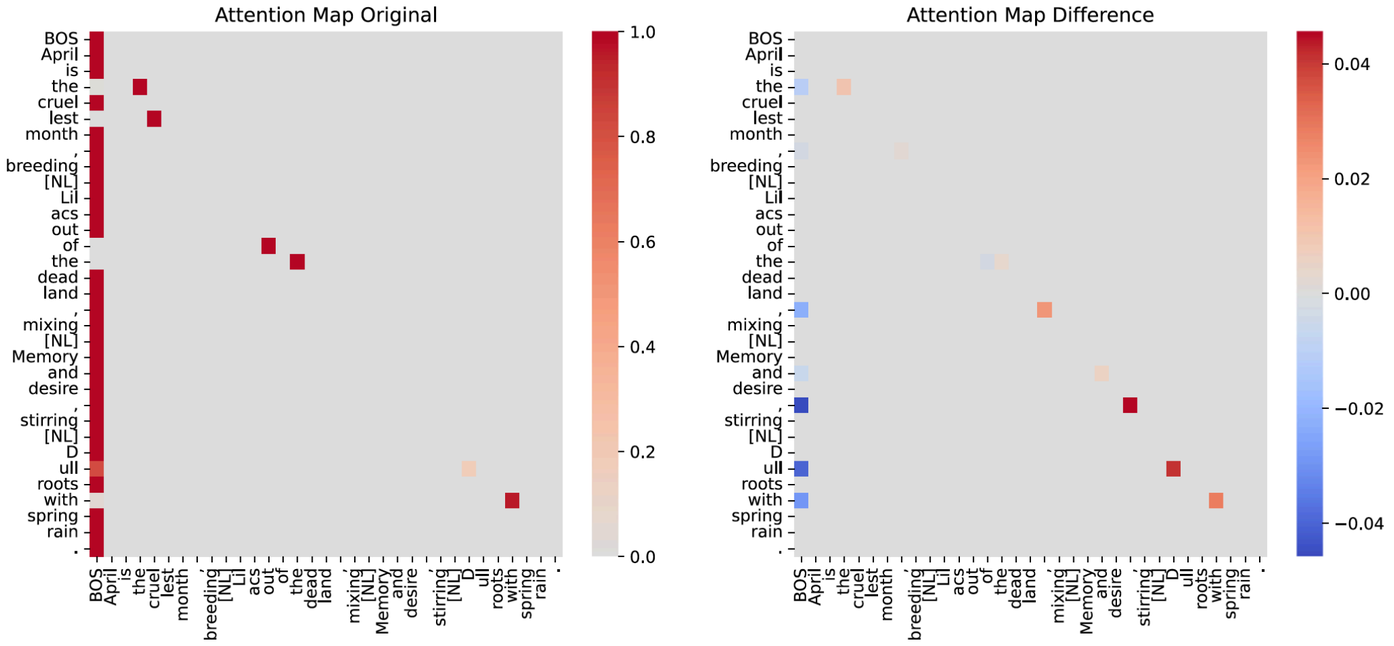


Fig. 11 | Example of attention sink shift from Llama 3-8B layer 0 head 6. The example sentence is the first several lines of T. S. Eliot's long poem *The Waste Land*. Note that for visualization purposes, the attention values are not divided by the scaling factor before the softmax operation.

Findings

- **Findings 3**
  - Perturbing ToM-sensitive parameter patterns affects the **attention mechanism**, thereby influencing language understanding.
  - Perturbing the ToM-sensitive parameter pattern alters the angle between q and BOS k under **positional encoding**.
  - This disruption breaks the **RoPE encoding**, causing q and BOS k to become more orthogonal.
  - As a result, the **attention sink is destabilized**, distorting the attention matrix and impairing the model's ability to capture correct feature relationships, ultimately diminishing its ToM capabilities.

Conclusion I: The Structural Fragility

- RQ1: How do LLMs encode Theory-of-Mind capabilities?
  - (Focus: Internal Mechanisms & Sparsity)

- **Sparse & Fragile Encoding**: ToM capabilities are not distributed evenly but are encoded in extremely sparse parameter patterns.
- **Mechanism of Failure**: This fragility stems from a deep dependency on Positional Encoding. Perturbing these parameters disrupts dominant frequencies, severing the link between tokens and their positions.
- **The "Attention Sink" Effect**: The structural breakdown leads to a collapse in Attention Sinks, causing the model to lose contextual focus and fail at reasoning tasks.

Research Questions

- **RQ1**: How do LLMs encode Theory-of-Mind capabilities?
  - (Focus: Internal Mechanisms & Sparsity)

- **RQ2**: How can we enhance Theory-of-Mind reasoning in LLMs?
  - (Focus: Inference-Time Scaling)

How can we enhance ToM in LLMs?

- **Step 1:** Formalization via Dynamic Epistemic Logic (DEL)
  - We reframe ToM reasoning not as text generation, but as a structured dynamic belief update process grounded in DEL.
- **Step 2:** Process-Level Supervision
  - We train a Process Belief Model (PBM) using noise-free, symbolic labels generated by a DEL simulator to verify the correctness of each intermediate reasoning step.
- **Step 3:** Inference-Time Scaling
  - We apply PBM-guided search strategies (e.g., Best-of-N, Beam Search) to select reliable reasoning traces, allocating test-time compute to boost performance without retraining.

Formalizing ToM Based on Modal Logic

- **The Core Idea: Reasoning as Calculation**
  - Instead of treating ToM as abstract "intuition," we model it using DEL. This gives us a calculable "World Model."
- **How It Works**
  - **States** (Worlds): Represent all possibilities of reality (e.g., Chocolate is in Drawer vs. Table).
  - **Beliefs**: Defined by which worlds an agent considers "possible."
  - **Updates**: Every action (e.g., "John leaves the room") is a function that updates these possibilities.

Formalizing ToM Based on Modal Logic

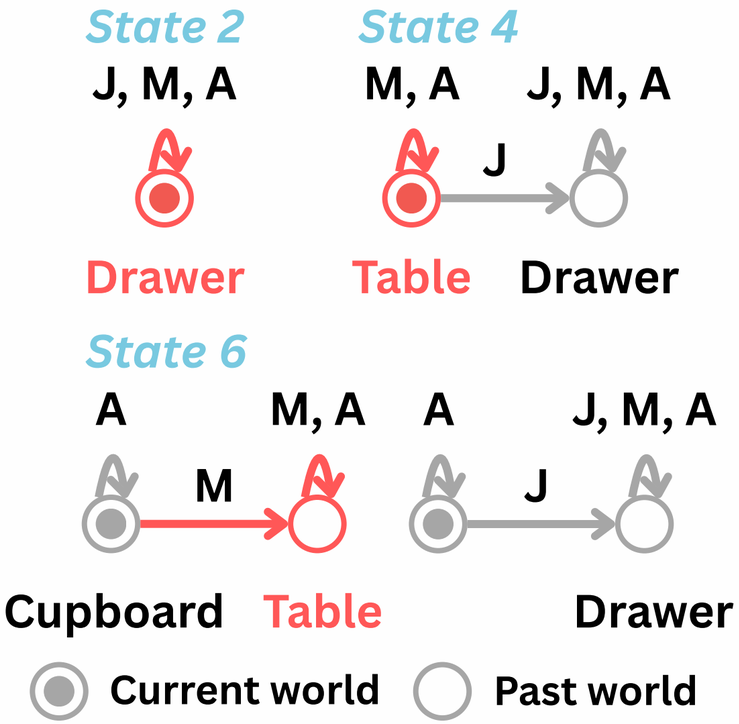
ToM Actions

1 John (J), Mary (M) and Alice (A) entered the kitchen.
2 John put the chocolate in the **drawer**. *State 2*
3 John exited the kitchen.
4 Mary moved the chocolate to the **table**. *State 4*
5 Mary exited the kitchen.
6 Alice moved the chocolate to the **cupboard**. *State 6*
Question: Where does **Mary** think **Alice** thinks the chocolate is?

Process Label Generation

State 1	*State 2*	*State 3*	*State 4*	*State 5*	*State 6*
Null	Drawer	Drawer	Table	Table	Table

DEL Belief States

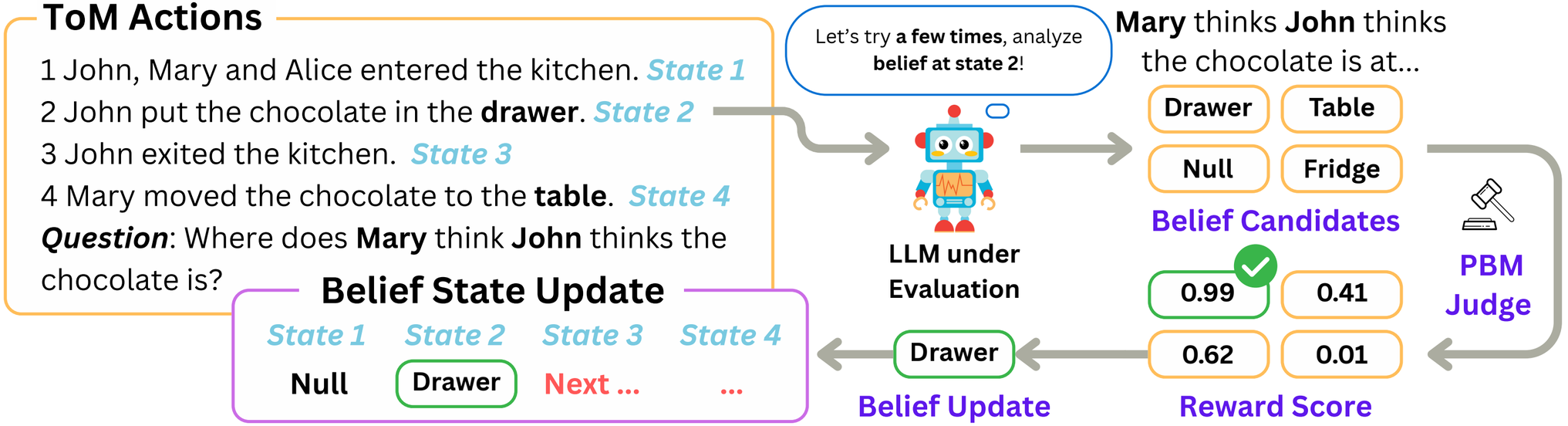


Building the Verifier (PBM)

- **Synthetic Data Generation:**
  - We integrated a DEL Simulator to automatically synthesize 20,000 ToM stories.
  - The "Gold" Standard: Unlike human annotation, the simulator provides noise-free, symbolic labels for every single belief update step under DEL.
- **Training the Process Belief Model (PBM):**
  - Input: A reasoning step generated by an LLM (e.g., GPT-4o-mini).
  - Task: The PBM acts as a binary classifier (Judge), scoring each step as "Correct" or "Incorrect" based on the DEL ground truth.
  - Outcome: A lightweight verifier that learns the "logic of belief" without needing expensive retraining of the base model.

Test-Time Scaling for ToM

- **The Pipeline**: Generate, Verify, Select
  - **Generator**: The LLM proposes multiple candidate reasoning traces (belief updates) in parallel.
  - **Verifier**: The PBM scores each step/trace, acting as a reward model.
  - **Selector**: We use search algorithms (e.g., Best-of-N, Beam Search) to identify the most reliable path based on PBM scores.



Main Results

Table 1: Inference-time scaling across belief orders in the Hi-ToM dataset using BoN and Beam Search. "Ori" denotes baseline accuracy, and "+PBM" denotes accuracy with inference-time scaling.

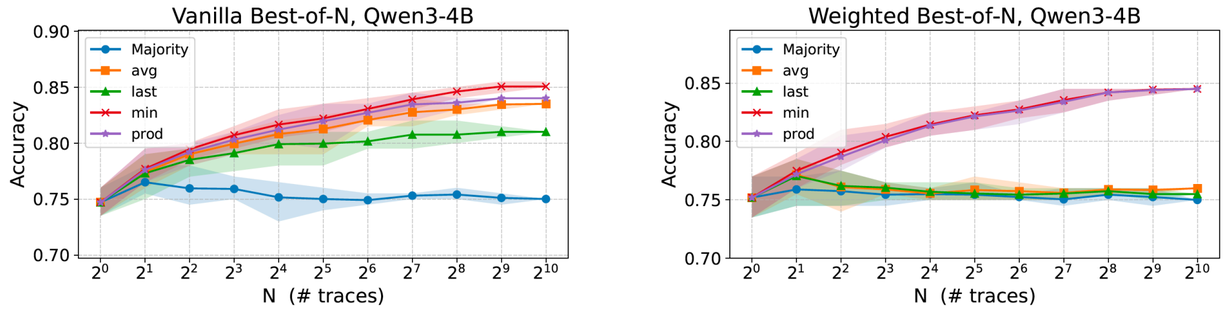
Model	0-th Order		1-th Order		2-th Order		3-th Order		4-th Order		Average	
	Ori	+PBM	Ori	+PBM	Ori	+PBM	Ori	+PBM	Ori	+PBM	Ori	+PBM
BoN ($N = 1024$)												
Qwen3-4B	100.0	100.0	79.8	85.0	79.3	90.0	70.2	82.5	46.0	65.0	75.1	84.5
Qwen3-1.7B	78.0	82.5	59.7	65.0	45.2	55.0	47.0	62.5	47.8	57.5	55.5	64.5
Qwen3-0.6B	69.2	80.0	52.0	72.5	35.0	47.5	31.5	52.5	34.0	47.5	44.3	60.0
Llama3.2-3B	68.2	85.0	52.0	80.0	43.2	82.5	37.0	82.5	36.8	75.0	47.4	81.0
Llama3.2-1B	41.5	46.2	40.0	53.8	28.5	61.5	41.5	84.6	29.2	58.3	36.1	60.9
BoN ($N = 4$)												
gpt-4.1	95.0	97.5	85.0	87.5	85.0	92.5	82.5	95.0	70.0	77.5	83.5	90.0
gpt-4.1-mini	77.5	70.0	90.0	85.0	70.0	75.0	75.0	92.5	77.5	92.5	78.0	83.0
gpt-4o	100.0	100.0	85.0	90.0	82.5	92.5	90.0	97.5	77.5	85.0	87.0	93.0
gpt-4o-mini	90.0	100.0	75.0	87.5	77.5	95.0	77.5	100.0	55.0	85.0	75.0	93.5
Beam Search ($N = 256$)												
Qwen3-8B	96.5	80.0	53.3	80.0	38.8	85.0	55.8	95.0	57.8	95.0	60.4	87.0
Qwen3-4B	100.0	100.0	79.8	85.0	79.3	97.5	70.2	82.5	46.0	60.0	75.1	85.0

SOTA Comparison & Generalization

- **Closing the Gap**: Small Models > Giants
- **Out-of-Distribution Generalization**: Beyond Synthetic Data

Table 2: Comparison with SOTA LLMs on Hi-ToM (BoN, $N = 1024$). "+PBM" denotes accuracy with inference-time scaling.

Model	0-th	1-th	2-th	3-th	4-th	Avg.
o4-mini	97.5	95.0	77.5	87.5	85.0	88.5
gpt-4o	100.0	85.0	82.5	90.0	77.5	87.0
Qwen3-4B+PBM	**100.0**	**85.0**	**90.0**	**82.5**	**65.0**	**84.5**
Qwen3-235B-A22B	100.0	75.0	85.0	85.0	75.0	84.0
gpt-4.1	95.0	85.0	85.0	82.5	70.0	83.5
DeepSeek-V3	100.0	80.0	90.0	70.0	72.5	82.5
Llama3.2-3B+PBM	**85.0**	**80.0**	**82.5**	**82.5**	**75.0**	**81.0**
gpt-4.1-mini	77.5	90.0	70.0	75.0	77.5	78.0
gpt-4o-mini	90.0	75.0	77.5	77.5	55.0	75.0
Qwen3-1.7B+PBM	**82.5**	**65.0**	**55.0**	**62.5**	**57.5**	**64.5**
OLMo-32B	77.5	60.0	60.0	65.0	52.5	63.0
Llama3.2-1B+PBM	**46.2**	**53.8**	**61.5**	**84.6**	**58.3**	**60.9**
Qwen3-0.6B+PBM	**80.0**	**72.5**	**47.5**	**52.5**	**47.5**	**60.0**
gpt-4.1-nano	22.5	32.5	42.5	27.5	30.0	31.0



(a) Vanilla BoN decoding on Qwen3-4B.

(b) Weighted BoN decoding on Qwen3-4B.

Figure 3: Accuracy of BoN decoding on Qwen3-4B across different budgets N in the Hi-ToM dataset. Results are shown for (a) Vanilla and (b) Weighted aggregation strategies.

Table 3: BoN ($N = 1024$) inference-time scaling on the dataset from Kosinski (Kosinski, 2024), evaluated across different belief types. "Ori" denotes baseline accuracy; "+PBM" denotes accuracy with inference-time scaling.

Model	False Belief		Informed Protagonist		No Transfer		Present Protagonist		Average	
	Ori	+PBM	Ori	+PBM	Ori	+PBM	Ori	+PBM	Ori	+PBM
Qwen3-8B	83.3	87.5	83.8	85.0	92.8	97.5	79.5	85.0	84.8	88.8
Qwen3-4B	70.2	80.0	86.2	90.0	93.2	95.0	88.0	92.5	84.4	89.4
Qwen3-1.7B	18.2	35.0	15.5	37.5	24.8	60.0	13.8	30.0	18.1	40.6
Qwen3-0.6B	14.5	12.5	23.5	30.0	25.0	35.0	21.0	32.5	21.0	27.5

Analysis: Reliability, Scalability & Cost

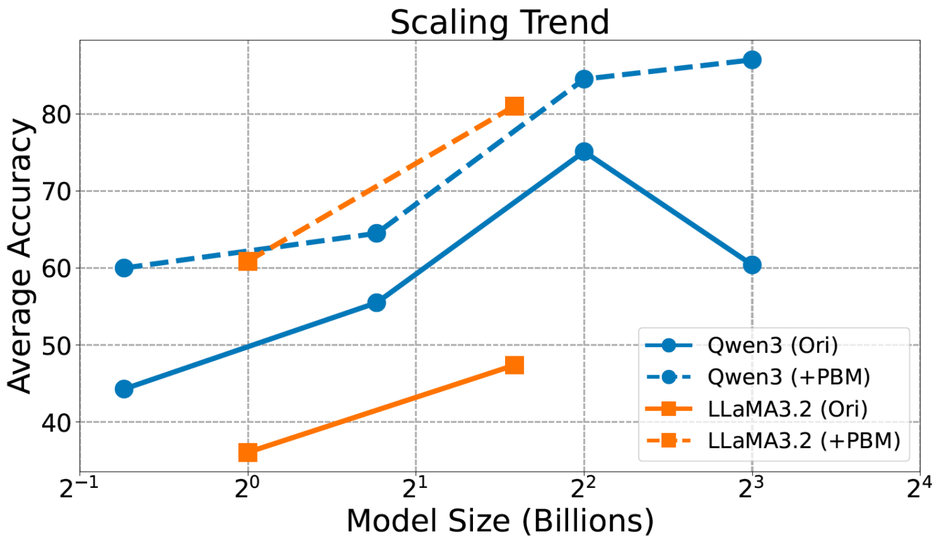
- **Benchmarking the Verifier (PBM Quality)**
- **Scaling with Model Size**
- **Cost Efficiency**

Table 4: PBM classification accuracy (%) across belief orders on the test set.

PBM	0-th	1-th	2-th	3-th	4-th	Avg.
Llama3.1-8B	99.2	94.6	89.0	87.0	79.9	90.0
Llama3.2-3B	99.1	91.9	84.9	83.8	73.8	86.7

Table 5: BoN inference-time scaling accuracy (%) on Hi-ToM using different PBMs.

Model+PBM	0-th	1-th	2-th	3-th	4-th	Avg.
Qwen3-4B + 8B	100.0	85.0	90.0	82.5	65.0	84.5
Qwen3-4B + 3B	100.0	77.5	77.5	72.5	47.5	75.0
Qwen3-1.7B + 8B	82.5	65.0	55.0	62.5	57.5	64.5
Qwen3-1.7B + 3B	82.5	60.0	45.0	47.5	50.0	57.0
Qwen3-0.6B + 8B	80.0	72.5	47.5	52.5	47.5	60.0
Qwen3-0.6B + 3B	77.5	55.0	27.5	35.0	32.5	45.5



Model	Input	Cached Input	Output	Total
gpt-4.1	$2.00	$0.50	$8.00	$10.50
gpt-4.1-mini	$0.40	$0.10	$1.60	$2.10
gpt-4o	$2.50	$1.25	$10.00	$13.75
gpt-4o-mini	$0.15	$0.075	$0.60	$0.825

Conclusion II: Enhancing & Scaling ToM

- RQ2: How can we enhance Theory-of-Mind reasoning in LLMs?
  - (Focus: Inference-Time Scaling)

- **ToM as Dynamic Logic**: We proposed DEL-TOM, a framework that reframes social reasoning as a structured, verifiable sequence of belief updates grounded in formal modal logic.
- **The Power of Test-Time Scaling**: By training a Process Belief Model (PBM) to verify these updates, we demonstrated that scaling test-time compute is a highly effective strategy, allowing smaller models to outperform larger ones without retraining.

Final Takeaways & Thoughts

- **Structural Fragility & Compression Risks**
  - ToM reasoning relies on extremely sparse parameter patterns.
  - Caution: Be careful with model compression! Quantization or pruning risks inadvertently wiping out these critical "social circuits."

Final Takeaways & Thoughts

- **Structural Fragility & Compression Risks**
  - ToM reasoning relies on extremely sparse parameter patterns.
  - Caution: Be careful with model compression! Quantization or pruning risks inadvertently wiping out these critical "social circuits."
- **The Shift to Generation-Verification**
  - To overcome this intrinsic fragility, we cannot rely solely on the base model's weights.
  - Solution: We can move toward a Generation-Verification pipeline (Test-Time Scaling) to filter out hallucinations and ensure reliability.
- **Formal Methods for trustworthy AI**
  - Verification is the core bottleneck. Formal Methods (like Dynamic Epistemic Logic) are essential.

Final Takeaways & Thoughts

- **Wait, is that intelligence? Is test-time scaling good?**
- 24 years ago..

Verification, The Key to AI

by Rich Sutton

November 15, 2001

It is a bit unseemly for an AI researcher to claim to have a special insight or plan for how his field should proceed. If analyzing the field as a whole, for diagnosing the ills that repeatedly plague it, and to suggest general solutions.

The insight that I would claim to have is that the key to a successful AI is that it can tell for itself whether or not it is assessment and make any necessary modifications. An AI that can assess itself may be able to make the modification

The Verification Principle:

> An AI system can create and maintain knowledge only to the extent that it can verify that knowledge itself.

Well, we don't even know if humans are truly reasoning or just doing advanced pattern matching. Hopefully, one day, we'll achieve Self-Generating, Self-Verifying models.

Thanks

- **How large language models encode theory-of-mind: a study on sparse parameter patterns**
  - Yuheng Wu, Wentao Guo, Zirui Liu, Heng Ji, Zhaozhuo Xu, Denghui Zhang
  - npj Artificial Intelligence 2025
  - https://www.nature.com/articles/s44387-025-00031-9
- **DEL-ToM: Inference-Time Scaling for Theory-of-Mind Reasoning via Dynamic Epistemic Logic**
  - Yuheng Wu, Jianwen Xie, Denghui Zhang, Zhaozhuo Xu
  - EMNLP 2025
  - https://aclanthology.org/2025.emnlp-main.573