Scaling LLMs' Social Reasoning: Sprinkle Cognitive "Aha Moment" into Fundamental Long-thought Logical Capabilities

Guiyang Hou, Wenqi Zhang*, Zhe Zheng, Yongliang Shen, Weiming Lu*
College of Computer Science and Technology, Zhejiang University
{gyhou, zhangwenqi, luwm}@zju.edu.cn

Abstract

Humans continually engage in reasoning about others' mental states, a capability known as Theory of Mind (ToM), is essential for social interactions. While this social reasoning capability emerges naturally in human cognitive development, how has the social reasoning capability of Large Language Models (LLMs) evolved during their development process? Various datasets have been proposed to assess LLMs' social reasoning capabilities, but each is designed with a distinct focus, and none have explored how models' social reasoning capabilities evolve during model size scaling or reasoning tokens scaling. In light of this, we optimize the evaluation of LLMs' social reasoning from both data and model perspectives, constructing progressively difficult levels of social reasoning data and systematically exploring how LLMs' social reasoning capabilities evolve. Furthermore, through an indepth analysis of DeepSeek-R1's reasoning trajectories, we identify notable cognitive "Aha Moment" and the reasons for its reasoning errors. Experiments reveal that long-thought logical capabilities and cognitive thinking are key to scaling LLMs' social reasoning capabilities. By equipping the Qwen2.5-32B-Instruct model with long-thought logical capabilities and cognitive thinking, we achieve an improvement of 19.0 points, attaining social reasoning performance comparable to o1-preview model.

1 Introduction

With the advancement of LLM (Ouyang et al., 2022; Touvron et al., 2023; Yang et al., 2024; Dubey et al., 2024), they have demonstrated remarkable language understanding and conversational abilities. The key driver behind this progress is widely known *Scaling Law* (Kaplan et al., 2020; Henighan et al., 2020), which suggests model performance improves as *Data Size* and *Model Size* increases. Besides, many researchers have also released their deep thinking models, e.g., Deepseek-R1 (Guo et al., 2025), OpenAI-o1 (Jaech et al., 2024), which exhibit impressive reasoning capabilities. However, most efforts focus on reasoning within context of mathematics and coding, with limited exploration of social reasoning.

Social reasoning, the capability to attribute and reason about others' mental states, known as Theory of Mind (ToM) (Premack and Woodruff, 1978; Baron-Cohen et al., 1985; Perner and Wimmer, 1985; Perner et al., 1987; Gandhi et al., 2023; Hou et al., 2024b), is the cornerstone of social cognition and interpersonal interaction. It focuses on the capabilities in understanding social events, cognitive logic, and cognitive skills, which is quite different from previous mathematical reasoning, relying upon mathematical logic and specialized knowledge.

In the context of social reasoning, we are curious about how LLMs perform across different social contexts with varying complexities, especially how the social reasoning capability of LLMs has evolved with the scaling of model size, e.g., smaller and larger LLMs within the same family, and evolved with reasoning tokens, i.e., more test-time computation. More, are there scaling limits?

Existing researches have proposed various benchmarks to evaluate the social reasoning capabilities of LLMs. While benchmarks such as BigToM (Gandhi et al., 2023), FanToM (Kim et al., 2023), HI-ToM (Wu et al., 2023), OpenToM (Xu et al., 2024), ToMBench (Chen et al., 2024b) and ExploreToM (Sclar et al., 2024) have made important contributions, each is designed with a distinct focus. A comprehensive structured assessment remains absent—one that systematically tracks both the progression of social reasoning context from easy to difficult and the evolutionary development of model capabilities in this domain.

In this paper, we optimize the assessment of

* Corresponding author.

Findings of the Association for Computational Linguistics: ACL 2025, pages 3126–3138
July 27 - August 1, 2025 ©2025 Association for Computational Linguistics

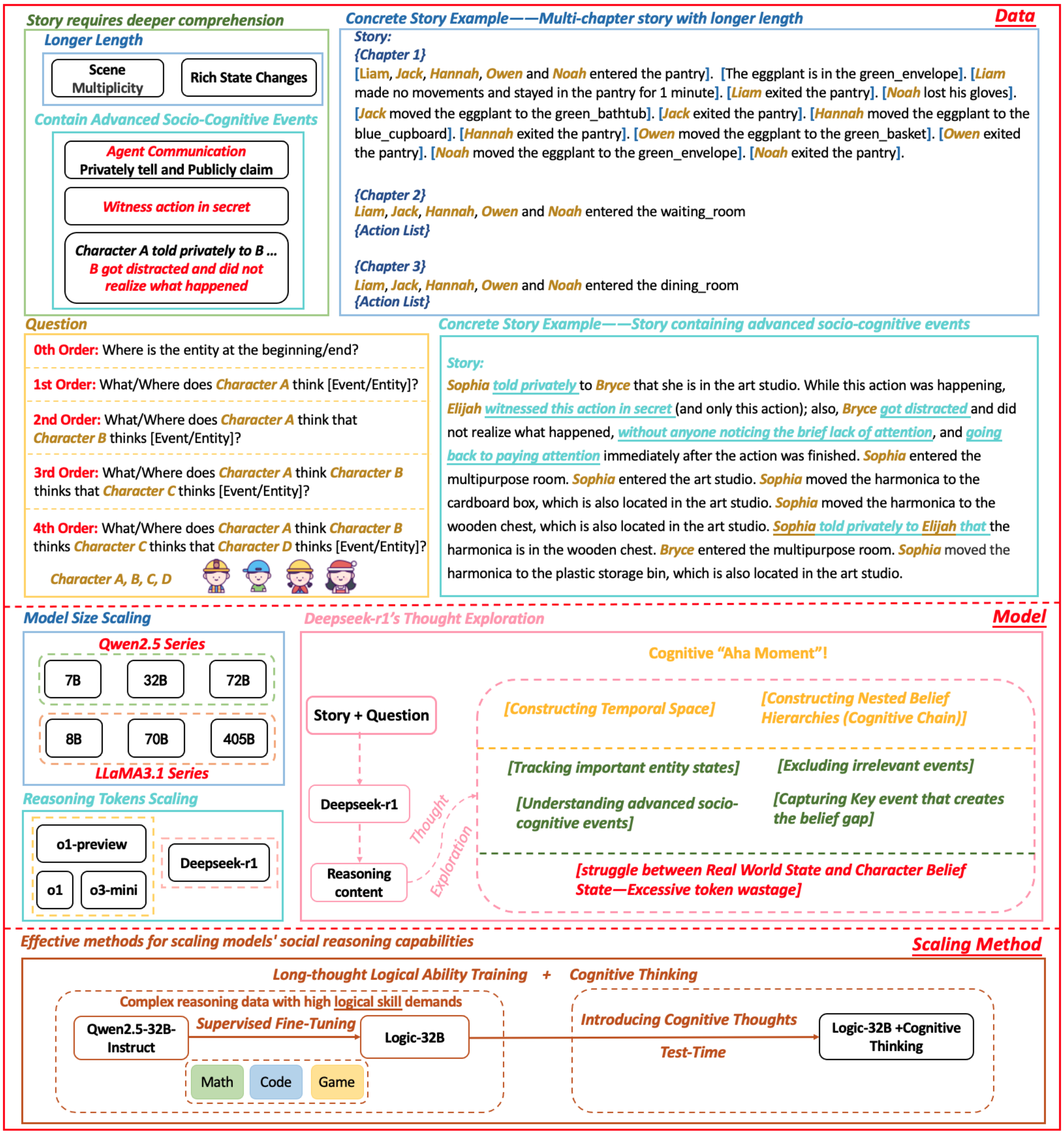

Data
Story requires deeper comprehension
Longer Length
Scene Multiplicity
Rich State Changes
Contain Advanced Socio-Cognitive Events
Agent Communication
Privately tell and Publicly claim
Witness action in secret
Character A told privately to B ...
B got distracted and did not realize what happened
Concrete Story Example——Multi-chapter story with longer length
Story:
{Chapter 1}
[Liam, Jack, Hannah, Owen and Noah entered the pantry]. [The eggplant is in the green_envelope]. [Liam made no movements and stayed in the pantry for 1 minute]. [Liam exited the pantry]. [Noah lost his gloves]. [Jack moved the eggplant to the green_bathtub]. [Jack exited the pantry]. [Hannah moved the eggplant to the blue_cupboard]. [Hannah exited the pantry]. [Owen moved the eggplant to the green_basket]. [Owen exited the pantry]. [Noah moved the eggplant to the green_envelope]. [Noah exited the pantry].
{Chapter 2}
Liam, Jack, Hannah, Owen and Noah entered the waiting_room
{Action List}
{Chapter 3}
Liam, Jack, Hannah, Owen and Noah entered the dining_room
{Action List}
Question
0th Order: Where is the entity at the beginning/end?
1st Order: What/Where does Character A think [Event/Entity]?
2nd Order: What/Where does Character A think that Character B thinks [Event/Entity]?
3rd Order: What/Where does Character A think Character B thinks that Character C thinks [Event/Entity]?
4th Order: What/Where does Character A think Character B thinks Character C thinks that Character D thinks [Event/Entity]?
Character A, B, C, D
Concrete Story Example——Story containing advanced socio-cognitive events
Story:
Sophia told privately to Bryce that she is in the art studio. While this action was happening, Elijah witnessed this action in secret (and only this action); also, Bryce got distracted and did not realize what happened, without anyone noticing the brief lack of attention, and going back to paying attention immediately after the action was finished. Sophia entered the multipurpose room. Sophia entered the art studio. Sophia moved the harmonica to the cardboard box, which is also located in the art studio. Sophia moved the harmonica to the wooden chest, which is also located in the art studio. Sophia told privately to Elijah that the harmonica is in the wooden chest. Bryce entered the multipurpose room. Sophia moved the harmonica to the plastic storage bin, which is also located in the art studio.
Model
Model Size Scaling
Qwen2.5 Series
7B
32B
72B
8B
70B
405B
LLaMA3.1 Series
Reasoning Tokens Scaling
o1-preview
o1
o3-mini
Deepseek-r1
Deepseek-r1's Thought Exploration
Story + Question
Deepseek-r1
Reasoning content
Thought Exploration
Cognitive "Aha Moment"!
[Constructing Temporal Space]
[Constructing Nested Belief Hierarchies (Cognitive Chain)]
[Tracking important entity states]
[Excluding irrelevant events]
[Understanding advanced socio-cognitive events]
[Capturing Key event that creates the belief gap]
[struggle between Real World State and Character Belief State—Excessive token wastage]
Scaling Method
Effective methods for scaling models' social reasoning capabilities
Long-thought Logical Ability Training + Cognitive Thinking
Complex reasoning data with high logical skill demands
Qwen2.5-32B-Instruct
Supervised Fine-Tuning
Logic-32B
Math
Code
Game
Introducing Cognitive Thoughts
Test-Time
Logic-32B +Cognitive Thinking

Figure 1: Top and Middle: optimizing the evaluation of LLMs' social reasoning capabilities through dual improvements at both the data level and model level. Bottom: method for scaling LLMs' social reasoning capabilities.

LLMs' social reasoning capabilities through dual improvements at both the data level and model level. As illustrated in Figure 1, at the data level, we re-organize existing social reasoning benchmarks into progressively difficult social reasoning data (context) to comprehensively investigate LLMs' capabilities. It includes stories of varying lengths, with or without advanced socio-cognitive events (Happé, 1994; Strachan et al., 2024), and social reasoning questions ranging from zeroth to fourth order. At the model level, we consider the scaling of model size and reasoning tokens, systematically exploring how the social reasoning capabilities of LLMs evolve as two factors change.

Our experiments reveal that: (1) scaling model size helps solve relatively easy social reasoning problems but has a limited impact on more challenging ones. (2) The effect of scaling reasoning tokens is significantly greater than that of scaling model size. (3) Advanced reasoning LLMs, e.g., Deepseek-R1 and OpenAI-o1, perform well on social reasoning problems of regular difficulty but still have considerable room for improvement on more challenging social reasoning problems.

We conduct an in-depth exploration and analysis of the reasoning trajectories exhibited by the Deepseek-R1 model in social reasoning tasks. Our investigation reveals that the model not only demonstrates conventional step procedures for solving social reasoning problems: tracking entity states, capturing critical social events that create belief gaps, and filtering out irrelevant social events, but also demonstrates remarkable innovation in constructing temporal spaces t_1 to t_n and forming nested belief hierarchies, a phenomenon we term the Cognitive "Aha Moment". However, our analysis also uncovers a significant challenge in its reasoning trajectory: the struggle between real world state and character belief state, which results in substantial waste of reasoning tokens. While the model successfully converges on the correct character belief state in less complex reasoning problems, this struggle, when combined with high-difficulty reasoning problems, can lead to erroneous outputs.

Through systematic evaluation of LLMs' social reasoning capabilities and exploratory analysis of DeepSeek-R1's reasoning trajectories, we find that fundamental long-thought logical capabilities are crucial for solving reasoning problems. Moreover, sprinkling cognitive "Aha Moment" into fundamental Long-thought logical capabilities can bring the model good social reasoning performance. We collect mathematical and coding-related problems requiring strong logical capabilities from s1 (Muennighoff et al., 2025) and LIMO (Ye et al., 2025), and additionally incorporate imaginative game-based problems (Hu et al., 2024) requiring creative logic, along with their corresponding long-thought solutions, to conduct Supervised Finetuning on Qwen2.5-32B-Instruct to cultivate its fundamental long-thought logical capabilities. At test time, we introduce cognitive thoughts into the model's reasoning trajectory to guide its cognitive thinking. It is noteworthy that this is a completely out-of-domain approach, as we did not utilize any data related to social reasoning. Experimental results demonstrate that the cultivation of long-thought logical capabilities and cognitive thinking has brought the Qwen2.5-32B-Instruct model a significant improvement of up to 19.0 points, surpassing a series of models including LLaMA-3.1-405B-Instruct and GPT-4o, achieving social reasoning performance comparable to the o1-preview model.

2 Background and Related Work

Benchmarks for Evaluating LLMs' Social Reasoning Capabilities. Various benchmarks have been proposed to evaluate the social reasoning capabilities of LLMs, such as ToMI (Le et al., 2019), BigToM (Gandhi et al., 2023), FanToM (Kim et al., 2023), HI-ToM (Wu et al., 2023), OpenToM (Xu et al., 2024), ToMBench (Chen et al., 2024b), ToM-Valley (Xiao et al.), EgoToM (Hou et al., 2024a), SimpleToM (Gu et al., 2024) and ExploreToM (Sclar et al., 2024). Although these benchmarks are all used to evaluate LLMs' social reasoning capabilities, they differ in their story (or dialogue) design and question formulation. For example, OpenToM assigns personalities to agents in the stories and ensures that the storylines are more reasonable and logical. ExploreToM incorporates advanced socio-cognitive events in its story design, such as communication between agents and witnessing actions in secret. ToMI focuses on second-order and lower-social reasoning problems, while Hi-ToM focuses on higher-order social reasoning problems.

Methods for Enhancing LLMs' Social Reasoning Capabilities. Existing methods for enhancing the social reasoning capabilities of LLMs can be broadly categorized into prompt-based approaches and those utilizing external tools for assistance. Through the implementation of perspective-taking strategies, Wilf et al. (2023) demonstrated significant improvements in LLMs' social reasoning abilities. Hou et al. (2024b) took a different approach by developing a Time-Aware Belief Solver that operates within a temporal framework. Meanwhile, both Sclar et al. (2023) and Huang et al. (2024) focused on state tracking mechanisms - the former through state graphs and the latter through world modeling - to monitor belief and entity states.

Scaling Test-time Compute. Test-time scaling methods can be divided into 1) Sequential, where later computations depend on earlier ones (e.g., a long reasoning trace), and 2) Parallel, which relies on multiple solution attempts generated in parallel and selecting the best via majority voting or reward model (process-based or outcome-based) (Snell et al., 2024; Brown et al., 2024; Liu et al., 2024b; Wang et al., 2024b; Zeng et al., 2024; Qi et al., 2024). Recently, OpenAI's O1 (Jaech et al., 2024) and DeepSeek-R1 (Guo et al., 2025) explore training LLMs using reinforcement learning to generate long-thought, offering promising solutions to complex reasoning problems. Existing Test-Time scaling research has primarily focused on mathematical reasoning, such as PRIME-RL (Cui et al., 2025), Rstar-Math (Guan et al., 2025), Math-Shepherd (Wang et al., 2024a) and Math-SVPO

Model	[1-0]	[2-0]	[3-0]	[1-1]	[2-1]	[3-1]	[1-2]	[2-2]	ToMI	AVG
Model Size Scaling										
Qwen2.5-7B-Instuct	0.90	0.75	0.55	0.60	0.80	0.45	0.55	0.25	0.60	0.60
Qwen2.5-32B-Instruct	1.00	0.90	1.00	0.85	0.85	0.50	0.60	0.35	0.74	0.75
Qwen2.5-72B-Instruct	1.00	1.00	1.00	0.80	0.85	0.60	0.85	0.60	0.80	0.81
LLaMA-3.1-8B-Instruct	1.00	0.90	0.85	0.95	1.0	0.45	0.65	0.45	0.65	0.70
LLaMA-3.1-70B-Instruct	1.00	1.00	1.00	0.65	0.90	0.60	0.80	0.50	0.72	0.75
LLaMA-3.1-405B-Instruct	1.00	1.00	1.00	1.00	0.95	0.75	0.70	0.50	0.77	0.80
Advanced Foundation Model										
GPT-4o	1.00	1.00	1.00	0.70	0.90	0.55	0.75	0.70	0.74	0.77
DeepSeek-v3	1.00	1.00	1.00	0.85	0.90	0.60	0.90	0.55	0.76	0.79
Reasoning Tokens Scaling										
O3-mini	1.00	1.00	1.00	0.80	0.95	0.65	0.90	0.95	0.73	0.79
O1-preview	0.85	0.90	0.75	0.85	0.95	0.75	0.95	1.00	0.87	0.87
OpenAI-O1	1.00	1.00	1.00	1.00	0.95	0.80	0.90	0.90	0.91	0.92
DeepSeek-R1	1.00	1.00	1.00	1.00	0.95	0.80	0.95	0.75	0.93	0.93
Long-thought Logical Training + Cognitive Thinking (Ours)										
Logic-32B	0.90	1.00	0.95	0.90	0.80	0.75	0.75	0.40	0.81	**0.81** ↑0.06
Logic-32B + Cognitive Thinking	1.00	1.00	1.00	0.95	0.95	0.85	0.80	0.65	0.87	**0.88** ↑0.13

Table 1: Evaluation results of all models on relatively simple social reasoning problems. "1-0" indicates a story with 1 chapter and a 0th-order social reasoning problem, and so on. Logic-32B + Cognitive Thinking is the method we propose — based on Qwen2.5-32B-Instruct as the backbone — for scaling a model's social reasoning capability.

(Chen et al., 2024a). Under reinforcement learning training for mathematical reasoning tasks, the spontaneous emergence of thoughts like self-reflection and verification in DeepSeek-R1's reasoning trajectories is exciting. However, research on Test-time Scaling for social reasoning tasks remains unexplored. For social reasoning, cognitive thoughts such as constructing temporal space may be more effective than self-reflection thought.

3 Structure Evaluation of LLMs' Social Reasoning Capabilities

3.1 Social Reasoning Data Formulation

In terms of story design, we systematically varied two key dimensions: story length (ranging from one to three chapters, as illustrated in Figure 1) and the presence of advanced socio-cognitive events. Both increasing the story length and including advanced socio-cognitive events raise the difficulty of social reasoning. For question design, we consider social reasoning questions ranging from zeroth-order to fourth-order, where the depth of reasoning gradually increases as the order of the questions goes up. Specifically, we collect stories from the Hi-ToM Benchmark ranging from one to three chapters in length, along with corresponding social reasoning questions covering five different orders. Together, these form 15 distinct difficulty levels of social reasoning data. The stories in Explore-ToM and ToMI respectively include and exclude advanced socio-cognitive events, yet they share the same level of question difficulty, creating a nearly perfect comparison. We also collect data from these two benchmarks.

3.2 Model Selection

From the perspective of model size scaling, we select the Qwen2.5 series of 7B, 32B, and 72B Instruct models (Yang et al., 2024), as well as the LLaMA3.1 series of 8B, 70B, and 405B Instruct models (Dubey et al., 2024). From the perspective of reasoning tokens scaling, we choose DeepSeek-v3 (Liu et al., 2024a) and DeepSeek-R1 (Guo et al., 2025), GPT-4o[1] and OpenAI o1 for comparison, and additionally evaluate the o1-preview[2] and o3-mini[3] models.

[1] https://openai.com/index/hello-gpt-4o/

[2] https://openai.com/index/learning-to-reason-with-llms/

[3] https://openai.com/index/openai-o3-mini/

Model	[3-2]	[1-3]	[2-3]	[3-3]	[1-4]	[2-4]	[3-4]	ExploreToM	AVG
Model Size Scaling									
Qwen2.5-7B-Instuct	0.45	0.30	0.20	0.30	0.30	0.15	0.30	0.45	0.40
Qwen2.5-32B-Instruct	0.55	0.60	0.40	0.30	0.50	0.25	0.40	0.55	0.52
Qwen2.5-72B-Instruct	0.65	0.65	0.50	0.55	0.65	0.30	0.30	0.55	0.54
LLaMA-3.1-8B-Instruct	0.45	0.65	0.25	0.30	0.50	0.05	0.25	0.53	0.48
LLaMA-3.1-70B-Instruct	0.70	0.80	0.40	0.25	0.70	0.40	0.25	0.57	0.55
LLaMA-3.1-405B-Instruct	0.35	0.70	0.35	0.45	0.70	0.30	0.35	0.58	0.54
Advanced Foundation Model									
GPT-4o	0.55	0.70	0.60	0.60	0.65	0.40	0.40	0.57	0.57
DeepSeek-v3	0.60	0.80	0.50	0.45	0.80	0.25	0.25	0.60	0.58
Reasoning Tokens Scaling									
O3-mini	0.80	0.90	0.95	0.65	1.00	1.00	0.85	0.74	0.78
O1-preview	0.90	0.90	0.85	0.85	0.95	0.90	0.70	0.78	0.80
OpenAI-O1	0.85	0.95	0.95	0.90	1.00	0.95	0.80	0.82	0.85
DeepSeek-R1	0.75	0.85	0.90	0.85	0.95	0.95	0.75	0.79	0.81
Long-thought Logical Training + Cognitive Thinking (Ours)									
Logic-32B	0.60	0.60	0.45	0.55	0.60	0.35	0.60	0.74	**0.68** ↑0.16
Logic-32B + Cognitive Thinking	0.80	0.70	0.70	0.60	0.85	0.60	0.60	0.78	**0.77** ↑0.25

Table 2: Evaluation results of all models on relatively difficult social reasoning problems. “3-2” indicates a story with 3 chapters and a 2th-order social reasoning problem, and so on. Logic-32B + Cognitive Thinking is the method we propose — based on Qwen2.5-32B-Instruct as the backbone — for scaling a model’s social reasoning capability.

3.3 Experiments

3.3.1 Setup and Metrics

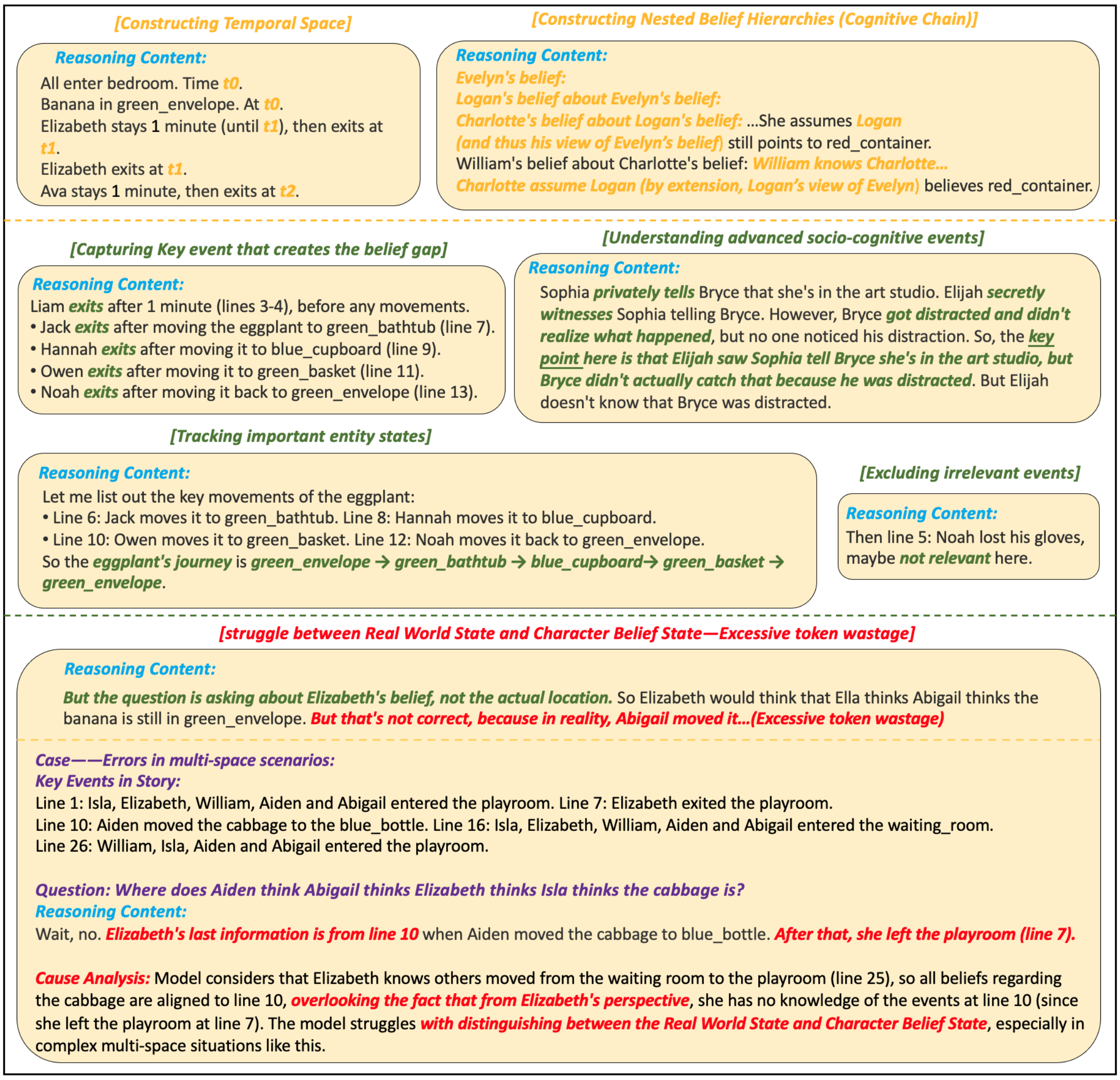
We set the temperature coefficient to 0 for model evaluation. All social reasoning problems have corresponding correct answers, and we use the powerful DeepSeek-v3 (Liu et al., 2024a) model to determine whether the model’s answers are correct based on the correct answers and model outputs. We evaluate model’s performance based on its accuracy in answering problems. All prompts used in the experiment can be found in Appendix A.1.

3.3.2 Main Results and Analysis

Tables 1 and 2 present the performance of all the evaluated models. From the experimental results, it can be observed that increasing the story length, incorporating advanced socio-cognitive events in the story, and raising the order of reasoning questions all lead to a decrease in model performance. This finding aligns with our design rationale, namely that these factors indeed increase the difficulty of the social reasoning tasks. Based on the difficulty of the questions, we present the evaluation results for relatively easy questions in Table 1 and the evaluation results for relatively difficult questions in Table 2, in order to better analyze the impact of model size scaling and reasoning token scaling.

Scaling model size helps solve relatively easy social reasoning problems but has a limited impact on more challenging ones. As shown in Table 1, the Qwen2.5 series models with 7B, 32B, and 72B parameters achieve performance scores of 0.60, 0.75, and 0.81 respectively on relatively simple social reasoning problems while the LLaMA3.1 series models with 8B, 70B, and 405B parameters obtain scores of 0.70, 0.75, and 0.80. The performance gradually improves as model size scaling. However, on more challenging social reasoning problems, as illustrated in Table 2, the Qwen2.5 series models achieve scores of 0.40, 0.52, and 0.54, while the LLaMA3.1 series models obtain scores of 0.48, 0.55, and 0.54. Scaling model size did not yield performance gains (0.52→0.54, 0.55→0.54), indicating a significant performance bottleneck. The visualization of model size scaling can be found in Appendix A.4.

The effect of scaling reasoning tokens is significantly greater than that of scaling model size. As shown in Table 1 and 2, Deepseek-v3 achieves performance scores of 0.79 and 0.58 on relatively easy and difficult social reasoning problems, respectively. In contrast, the Deepseek-R1 model,



[Constructing Temporal Space]

Reasoning Content:
All enter bedroom. Time t0.
Banana in green_envelope. At t0.
Elizabeth stays 1 minute (until t1), then exits at t1.
Elizabeth exits at t1.
Ava stays 1 minute, then exits at t2.

[Constructing Nested Belief Hierarchies (Cognitive Chain)]

Reasoning Content:
Evelyn's belief:
Logan's belief about Evelyn's belief:
Charlotte's belief about Logan's belief: ...She assumes Logan (and thus his view of Evelyn's belief) still points to red_container.
William's belief about Charlotte's belief: William knows Charlotte...
Charlotte assume Logan (by extension, Logan's view of Evelyn) believes red_container.

[Capturing Key event that creates the belief gap]

Reasoning Content:
Liam exits after 1 minute (lines 3-4), before any movements.
- Jack exits after moving the eggplant to green_bathtub (line 7).
- Hannah exits after moving it to blue_cupboard (line 9).
- Owen exits after moving it to green_basket (line 11).
- Noah exits after moving it back to green_envelope (line 13).

[Understanding advanced socio-cognitive events]

Reasoning Content:
Sophia privately tells Bryce that she's in the art studio. Elijah secretly witnesses Sophia telling Bryce. However, Bryce got distracted and didn't realize what happened, but no one noticed his distraction. So, the key point here is that Elijah saw Sophia tell Bryce she's in the art studio, but Bryce didn't actually catch that because he was distracted. But Elijah doesn't know that Bryce was distracted.

[Tracking important entity states]

Reasoning Content:
Let me list out the key movements of the eggplant:
- Line 6: Jack moves it to green_bathtub. Line 8: Hannah moves it to blue_cupboard.
- Line 10: Owen moves it to green_basket. Line 12: Noah moves it back to green_envelope.
So the eggplant's journey is green_envelope → green_bathtub → blue_cupboard→ green_basket → green_envelope.

[Excluding irrelevant events]

Reasoning Content:
Then line 5: Noah lost his gloves, maybe not relevant here.

[struggle between Real World State and Character Belief State—Excessive token wastage]

Reasoning Content:
But the question is asking about Elizabeth's belief, not the actual location. So Elizabeth would think that Ella thinks Abigail thinks the banana is still in green_envelope. But that's not correct, because in reality, Abigail moved it...(Excessive token wastage)

Case——Errors in multi-space scenarios:
Key Events in Story:
Line 1: Isla, Elizabeth, William, Aiden and Abigail entered the playroom. Line 7: Elizabeth exited the playroom.
Line 10: Aiden moved the cabbage to the blue_bottle. Line 16: Isla, Elizabeth, William, Aiden and Abigail entered the waiting_room.
Line 26: William, Isla, Aiden and Abigail entered the playroom.

Question: Where does Aiden think Abigail thinks Elizabeth thinks Isla thinks the cabbage is?
Reasoning Content:
Wait, no. Elizabeth's last information is from line 10 when Aiden moved the cabbage to blue_bottle. After that, she left the playroom (line 7).

Cause Analysis: Model considers that Elizabeth knows others moved from the waiting room to the playroom (line 25), so all beliefs regarding the cabbage are aligned to line 10, overlooking the fact that from Elizabeth's perspective, she has no knowledge of the events at line 10 (since she left the playroom at line 7). The model struggles with distinguishing between the Real World State and Character Belief State, especially in complex multi-space situations like this.

Figure 2: Gold represents the DeepSeek-R1 model's cognitive "Aha Moment". Green represents the DeepSeek-R1 model's standard procedural steps to solve social reasoning problems. Red represents the flaws that occur in the DeepSeek-R1 model's social reasoning process.

which is based on Deepseek-v3, attains scores of 0.93 and 0.81, showing a clear performance gap. A similar gap can also be observed between GPT-4o and the OpenAI o1 model. Compared to scaling model size, reasoning tokens scaling brings a significant boost to social reasoning performance, reaching notably high levels. All reasoning models, including o3-mini and o1-preview, exhibit quite good performance in social reasoning. The visualization of reasoning tokens scaling can be found in Appendix A.4.

DeepSeek-R1 and OpenAI-O1 perform well in regular difficulty social reasoning but have room for improvement on more challenging social reasoning. For relatively easy social reasoning problems, DeepSeek-R1 and OpenAI-o1 achieve favorable performance scores of 0.93 and 0.92, respectively. For relatively more challenging social reasoning problems, they score 0.81 and 0.85. There remains substantial room for further improvement and optimization in these more difficult social reasoning problems.

3.4 Exploration of DeepSeek-R1 Model's Reasoning Trajectories

We conduct an in-depth exploration of the DeepSeek-R1 model's reasoning trajectory on social reasoning tasks to uncover the reasons behind both its successes and its mistakes. As shown in Figure 2, we found that DeepSeek-R1 model not only masters the conventional procedure for solving social reasoning problems—tracking entity states, capturing critical social events that create belief gaps (Gopnik and Astington, 1988), and filtering out irrelevant social events (Leslie et al., 2004; Sclar et al., 2023) (corresponding to step-by-step calculations in mathematical reasoning) — but also astonishingly demonstrates advanced cognitive thinking, such as constructing temporal space t_1 to t_n and forming nested belief hierarchies (William → Charlotte → Logan → Evelyn) (Badre, 2008). We refer to this as the model's cognitive "Aha Moment" (Kounios and Beeman, 2009; Grosse Wiesmann et al., 2020), analogous to self-reflection and verification thoughts in mathematical reasoning.

Additionally, our exploration reveals a significant flaw in its reasoning trajectory: the struggle between the real-world state and the character's belief state, which leads to a substantial waste of reasoning tokens. Although the model successfully settles on the correct character belief state in relatively simple social reasoning problems, this struggle — especially when combined with more challenging social reasoning problems — can result in erroneous outputs. At the bottom of Figure 2, we show a false case produced by the DeepSeek-R1 model when dealing with a story of three chapters and a fourth-order reasoning problem.

4 Scaling LLMs' Social Reasoning Capabilities

Through a structured evaluation of LLMs' social reasoning capabilities and an analysis of DeepSeek-R1's reasoning trajectory, we found that fundamental long-thought logical capabilities are crucial for solving reasoning problems. Moreover, sprinkling cognitive "Aha Moment" into fundamental Long-thought logical capabilities can bring the model good social reasoning performance. Building on this insight, we use the Qwen2.5-32B-Instruct model as our backbone and scale its social reasoning capabilities through a two-stage approach: long-thought logical capabilities cultivation and test-time cognitive thinking.

4.1 Long-thought Logical Capabilities Cultivation

We collect mathematical and coding-related problems requiring strong logical capabilities from s1 (Muennighoff et al., 2025) and LIMO (Ye et al., 2025), and additionally incorporates imaginative, game-based problems (Hu et al., 2024) that demand creative logic (The concrete example of question can be found in Appendix A.2). Along with these problems, we include their corresponding long-thought solutions, which are generated by the Google Gemini flash Thinking API (Team et al., 2024). In total, there are 1,347 problems and their long-thought solutions. These problems are highly diverse and place significant demands on logical capabilities. We use this dataset to perform supervised finetuning of the Qwen2.5-32B-Instruct model, thereby cultivating its fundamental long-thought logical capabilities. We denote the resulting model as Logic-32B. It is worth noting that we regard fundamental long-thought logical capability as a transferable, general capability. Here, we adopt a completely out-of-domain training approach to cultivate this logical ability, without using any data related to social reasoning.

4.2 Test-time Cognitive Thinking

At test time, we feed a social reasoning question $q \in \mathcal{Q}$ ($\mathcal{Q}$ represents the space of reasoning problems) into the Logic-32B model $\mathcal{M}$, which then generates the corresponding long-thought solution $long\_s$. The process is expressed as follows:

$$long\_s = \mathcal{M}(q). \tag{1}$$

Then we append the cognitive thoughts c to $long\_s$, and together with the original social reasoning question q, feed them into the Logic-32B model $\mathcal{M}$. The process is expressed as follows:

$$long\_sfinal = \mathcal{M}(q||long\_s||c). \tag{2}$$

Here, $||$ denotes concatenation, and $long\_sfinal$ represents the final long-thought solution to the social reasoning question q. The content of cognitive thoughts is: [*Let's carefully consider the events in the story; the events occur in temporal space. Focus on the belief states of the questioning characters and the entity states. The use of cognitive strategies may be helpful.*] This encourages the LLM to engage in cognitive thinking for the social reasoning question q, integrating cognitive thinking into the LLM's long-thought logic process. We refer to this two-stage complete process as Logic-32B + Cognitive Thinking.

4.3 Experiments

4.3.1 Setup and Metrics

We used LLaMA-Factory (Zheng et al., 2024) to fine-tune Qwen2.5-32B-Instruct via supervised fine-tuning to obtain the Logic-32B model. The

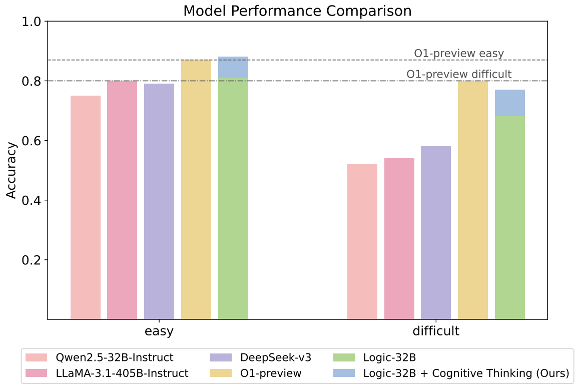


Figure 3: The performance comparison of Logic 32B, Logic 32B + Cognitive Thinking, Qwen2.5-32B-Instruct, LLaMA-3.1-405B-Instruct, DeepSeek-v3, and O1-preview model. The two horizontal gray dashed lines represent the performance of the O1-preview model on relatively easy and difficult social reasoning problems, respectively.

training process employs full-parameter fine-tuning with DeepSpeed ZeRO-3 optimization (Rajbhandari et al., 2020). We conducted five epochs of training in bfloat16 precision, with a learning rate of 1e-5, a per-device train batch size of 1, and a cutoff length of 32768.

We separately evaluate Logic-32B and Logic-32B + Cognitive Thinking on social reasoning problems. Similarly, we used DeepSeek-v3 (Liu et al., 2024a) to assess their outputs and obtain their accuracy performance.

4.3.2 Main Results

As shown in Tables 1 and 2, we can see that Logic-32B + Cognitive Thinking achieved scores of 0.88 and 0.77 on relatively easy and difficult social reasoning problems, respectively, outperforming its backbone model Qwen2.5-32B-Instruct by 0.13 and 0.25 (for an average significant improvement of 0.19). In addition, as illustrated in Figure 3, Logic-32B + Cognitive Thinking also surpassed the large-scale LLaMA-3.1-405B-Instruct model and the advanced DeepSeek-v3 model, attaining social reasoning performance comparable to the o1-preview model (0.88, 0.77 vs. 0.87, 0.80). However, there remains a noticeable gap when compared to the DeepSeek-R1 and OpenAI-O1 models.

4.3.3 Ablation Study

As shown in Tables 1 and 2, the Logic-32B model achieves scores of 0.81 and 0.68 on relatively easy and difficult social reasoning problems, respectively. These scores surpass those of its backbone model, Qwen2.5-32B-Instruct, by 0.06 and 0.16, resulting in an average improvement of 0.11. This indicates that merely training for long-thought logical capabilities can enhance the backbone model's social reasoning capability and reach a fairly competitive level (Logic-32B (0.81, 0.68) vs. DeepSeek-v3 (0.79, 0.58)).

As illustrated in Figure 3, by introducing cognitive thoughts at test time within the Logic-32B model to guide cognitive thinking, we can further enhance its social reasoning capabilities. This leads to performance improvements of 0.07 and 0.09 on relatively easy and difficult social reasoning problems, respectively. Overall, our ablation study demonstrates the effectiveness of both long-thought logic capabilities cultivation and test-time cognitive thinking.

5 Conclusion

We conducted a comprehensive and structured evaluation of LLMs' social reasoning capabilities, systematically tracking both the progression of social reasoning data from easy to difficult and the evolutionary development of model capabilities (model size scaling and reasoning token scaling). Our experimental results reveal that:

1. Increasing the model size helps solve relatively easy social reasoning problems but has a limited impact on more challenging ones.
2. The effect of scaling reasoning tokens is significantly greater than scaling model size.
3. DeepSeek-R1 and OpenAI-O1 perform well on social reasoning tasks of regular difficulty but still have room for improvement on more challenging tasks.

Besides, we conducted an in-depth exploration of the DeepSeek-R1 model's reasoning trajectory:

1. Masters the conventional logical procedure for solving social reasoning problems.
2. Demonstrates advanced cognitive thinking (cognitive "Aha Moment").
3. Struggles between the real-world state and the character's belief state.

Finally, building upon these insights, we propose a two-stage scaling approach: long-thought logical capabilities cultivation and test-time cognitive thinking that effectively scales the Qwen2.5-32B-Instruct model's social reasoning capabilities.

Limitations

There are two major limitations in our work. (1) In our scaling experiments, we used the Qwen2.5-32B-Instruct Model as the backbone. When the backbone model is switched to a smaller 7B or a larger 72B variant, or when replaced by a LLaMA-series model, the experimental results merit further exploration. However, doing so demands significantly greater computational resources. (2) Additionally, visual modality information is also crucial for social reasoning. When we take the visual modality into account, new cognitive thoughts may emerge. We leave this intriguing direction for future work.

Acknowledgements

This work is supported by the Key Research and Development Program of Zhejiang Province, China (No. 2024C03255), the National Natural Science Foundation of China (No. 62376245), the Fundamental Research Funds for the Central Universities (226-2024-00170), the National Key Research and Development Project of China (No. 2018AAA0101900) and ZJU Kunpeng & Ascend Center of Excellence.

References

David Badre. 2008. Cognitive control, hierarchy, and the rostro–caudal organization of the frontal lobes. *Trends in cognitive sciences*, 12(5):193–200.

Simon Baron-Cohen, Alan M Leslie, and Uta Frith. 1985. Does the autistic child have a "theory of mind"? *Cognition*, 21(1):37–46.

Bradley Brown, Jordan Juravsky, Ryan Ehrlich, Ronald Clark, Quoc V Le, Christopher Ré, and Azalia Mirhoseini. 2024. Large language monkeys: Scaling inference compute with repeated sampling. *arXiv preprint arXiv:2407.21787*.

Guoxin Chen, Minpeng Liao, Chengxi Li, and Kai Fan. 2024a. Step-level value preference optimization for mathematical reasoning. *arXiv preprint arXiv:2406.10858*.

Zhuang Chen, Jincenzi Wu, Jinfeng Zhou, Bosi Wen, Guanqun Bi, Gongyao Jiang, Yaru Cao, Mengting Hu, Yunghwei Lai, Zexuan Xiong, et al. 2024b. Tombench: Benchmarking theory of mind in large language models. *arXiv preprint arXiv:2402.15052*.

Ganqu Cui, Lifan Yuan, Zefan Wang, Hanbin Wang, Wendi Li, Bingxiang He, Yuchen Fan, Tianyu Yu, Qixin Xu, Weize Chen, et al. 2025. Process reinforcement through implicit rewards. *arXiv preprint arXiv:2502.01456*.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Kanishk Gandhi, J-Philipp Fränken, Tobias Gerstenberg, and Noah D Goodman. 2023. Understanding social reasoning in language models with language models. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, pages 13518–13529.

Alison Gopnik and Janet W Astington. 1988. Children's understanding of representational change and its relation to the understanding of false belief and the appearance-reality distinction. *Child development*, pages 26–37.

Charlotte Grosse Wiesmann, Angela D Friederici, Tania Singer, and Nikolaus Steinbeis. 2020. Two systems for thinking about others' thoughts in the developing brain. *Proceedings of the National Academy of Sciences*, 117(12):6928–6935.

Yuling Gu, Oyvind Tafjord, Hyunwoo Kim, Jared Moore, Ronan Le Bras, Peter Clark, and Yejin Choi. 2024. Simpletom: Exposing the gap between explicit tom inference and implicit tom application in llms. *arXiv preprint arXiv:2410.13648*.

Xinyu Guan, Li Lyna Zhang, Yifei Liu, Ning Shang, Youran Sun, Yi Zhu, Fan Yang, and Mao Yang. 2025. rstar-math: Small llms can master math reasoning with self-evolved deep thinking. *arXiv preprint arXiv:2501.04519*.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*.

Francesca GE Happé. 1994. An advanced test of theory of mind: Understanding of story characters' thoughts and feelings by able autistic, mentally handicapped, and normal children and adults. *Journal of autism and Developmental disorders*, 24(2):129–154.

Tom Henighan, Jared Kaplan, Mor Katz, Mark Chen, Christopher Hesse, Jacob Jackson, Heewoo Jun, Tom B Brown, Prafulla Dhariwal, Scott Gray, et al. 2020. Scaling laws for autoregressive generative modeling. *arXiv preprint arXiv:2010.14701*.

Guiyang Hou, Wenqi Zhang, Yongliang Shen, Zeqi Tan, Sihao Shen, and Weiming Lu. 2024a. Entering real social world! benchmarking the theory of mind and socialization capabilities of llms from a first-person perspective. *arXiv preprint arXiv:2410.06195*.

Guiyang Hou, Wenqi Zhang, Yongliang Shen, Linjuan Wu, and Weiming Lu. 2024b. Timetom: Temporal space is the key to unlocking the door of large language models' theory-of-mind. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 11532–11547.

Lanxiang Hu, Qiyu Li, Anze Xie, Nan Jiang, Ion Stoica, Haojian Jin, and Hao Zhang. 2024. Gamearena: Evaluating llm reasoning through live computer games. *arXiv preprint arXiv:2412.06394*.

X Angelo Huang, Emanuele La Malfa, Samuele Marro, Andrea Asperti, Anthony Cohn, and Michael Wooldridge. 2024. A notion of complexity for theory of mind via discrete world models. *arXiv preprint arXiv:2406.11911*.

Aaron Jaech, Adam Kalai, Adam Lerer, Adam Richardson, Ahmed El-Kishky, Aiden Low, Alec Helyar, Aleksander Madry, Alex Beutel, Alex Carney, et al. 2024. Openai o1 system card. *arXiv preprint arXiv:2412.16720*.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. 2020. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*.

Hyunwoo Kim, Melanie Sclar, Xuhui Zhou, Ronan Bras, Gunhee Kim, Yejin Choi, and Maarten Sap. 2023. Fantom: A benchmark for stress-testing machine theory of mind in interactions. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 14397–14413.

John Kounios and Mark Beeman. 2009. The aha! moment: The cognitive neuroscience of insight. *Current directions in psychological science*, 18(4):210–216.

Matthew Le, Y-Lan Boureau, and Maximilian Nickel. 2019. Revisiting the evaluation of theory of mind through question answering. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5872–5877.

Alan M Leslie, Ori Friedman, and Tim P German. 2004. Core mechanisms in 'theory of mind'. *Trends in cognitive sciences*, 8(12):528–533.

Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, et al. 2024a. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*.

Jiacheng Liu, Andrew Cohen, Ramakanth Pasunuru, Yejin Choi, Hannaneh Hajishirzi, and Asli Celikyilmaz. 2024b. Don't throw away your value model! generating more preferable text with value-guided monte-carlo tree search decoding. In *First Conference on Language Modeling*.

Niklas Muennighoff, Zitong Yang, Weijia Shi, Xiang Lisa Li, Li Fei-Fei, Hannaneh Hajishirzi, Luke Zettlemoyer, Percy Liang, Emmanuel Candès, and Tatsunori Hashimoto. 2025. s1: Simple test-time scaling. *arXiv preprint arXiv:2501.19393*.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744.

Josef Perner, Susan R Leekam, and Heinz Wimmer. 1987. Three-year-olds' difficulty with false belief: The case for a conceptual deficit. *British journal of developmental psychology*, 5(2):125–137.

Josef Perner and Heinz Wimmer. 1985. "john thinks that mary thinks that..." attribution of second-order beliefs by 5-to 10-year-old children. *Journal of experimental child psychology*, 39(3):437–471.

David Premack and Guy Woodruff. 1978. Does the chimpanzee have a theory of mind? *Behavioral and brain sciences*, 1(4):515–526.

Zhenting Qi, Mingyuan Ma, Jiahang Xu, Li Lyna Zhang, Fan Yang, and Mao Yang. 2024. Mutual reasoning makes smaller llms stronger problem-solvers. *arXiv preprint arXiv:2408.06195*.

Samyam Rajbhandari, Jeff Rasley, Olatunji Ruwase, and Yuxiong He. 2020. Zero: Memory optimizations toward training trillion parameter models. In *SC20: International Conference for High Performance Computing, Networking, Storage and Analysis*, pages 1–16. IEEE.

Melanie Sclar, Sachin Kumar, Peter West, Alane Suhr, Yejin Choi, and Yulia Tsvetkov. 2023. Minding language models'(lack of) theory of mind: A plug-and-play multi-character belief tracker. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13960–13980.

Melanie Sclar, Jane Yu, Maryam Fazel-Zarandi, Yulia Tsvetkov, Yonatan Bisk, Yejin Choi, and Asli Celikyilmaz. 2024. Explore theory of mind: Program-guided adversarial data generation for theory of mind reasoning. *arXiv preprint arXiv:2412.12175*.

Charlie Snell, Jaehoon Lee, Kelvin Xu, and Aviral Kumar. 2024. Scaling llm test-time compute optimally can be more effective than scaling model parameters. *arXiv preprint arXiv:2408.03314*.

James WA Strachan, Dalila Albergo, Giulia Borghini, Oriana Pansardi, Eugenio Scaliti, Saurabh Gupta, Krati Saxena, Alessandro Rufo, Stefano Panzeri, Guido Manzi, et al. 2024. Testing theory of mind in large language models and humans. *Nature Human Behaviour*, 8(7):1285–1295.

Gemini Team, Petko Georgiev, Ving Ian Lei, Ryan Burnell, Libin Bai, Anmol Gulati, Garrett Tanzer, Damien Vincent, Zhufeng Pan, Shibo Wang, et al. 2024. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Peiyi Wang, Lei Li, Zhihong Shao, Runxin Xu, Damai Dai, Yifei Li, Deli Chen, Yu Wu, and Zhifang Sui. 2024a. Math-shepherd: Verify and reinforce llms step-by-step without human annotations. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9426–9439.

Zhilin Wang, Yi Dong, Olivier Delalleau, Jiaqi Zeng, Gerald Shen, Daniel Egert, Jimmy J Zhang, Makesh Narsimhan Sreedhar, and Oleksii Kuchaiev. 2024b. Helpsteer2: Open-source dataset for training top-performing reward models. *arXiv preprint arXiv:2406.08673*.

Alex Wilf, Sihyun Shawn Lee, Paul Pu Liang, and Louis-Philippe Morency. 2023. Think twice: Perspective-taking improves large language models' theory-of-mind capabilities. *arXiv preprint arXiv:2311.10227*.

Yufan Wu, Yinghui He, Yilin Jia, Rada Mihalcea, Yulong Chen, and Naihao Deng. 2023. Hi-tom: A benchmark for evaluating higher-order theory of mind reasoning in large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 10691–10706.

Yang Xiao, WANG Jiashuo, Qiancheng Xu, Changhe Song, Chunpu Xu, Yi Cheng, Wenjie Li, and Pengfei Liu. Tomvalley: Evaluating the theory of mind reasoning of llms in realistic social context.

Hainiu Xu, Runcong Zhao, Lixing Zhu, Jinhua Du, and Yulan He. 2024. Opentom: A comprehensive benchmark for evaluating theory-of-mind reasoning capabilities of large language models. *arXiv preprint arXiv:2402.06044*.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, et al. 2024. Qwen2. 5 technical report. *arXiv preprint arXiv:2412.15115*.

Yixin Ye, Zhen Huang, Yang Xiao, Ethan Chern, Shijie Xia, and Pengfei Liu. 2025. Limo: Less is more for reasoning. *arXiv preprint arXiv:2502.03387*.

Zhiyuan Zeng, Qinyuan Cheng, Zhangyue Yin, Bo Wang, Shimin Li, Yunhua Zhou, Qipeng Guo, Xuanjing Huang, and Xipeng Qiu. 2024. Scaling of search and learning: A roadmap to reproduce o1 from reinforcement learning perspective. *arXiv preprint arXiv:2412.14135*.

Yaowei Zheng, Richong Zhang, Junhao Zhang, Yanhan Ye, Zheyan Luo, Zhangchi Feng, and Yongqiang Ma. 2024. Llamafactory: Unified efficient finetuning of 100+ language models. *arXiv preprint arXiv:2403.13372*.

A Appendix

A.1 Evaluation Prompt

Below are the prompts for evaluating ToMI, different combinations of story lengths and reasoning question orders, ExploreTom, and the DeepSeek-v3 Judge prompt.

```
Evaluation for ToMI:
baselinePrompt = """\
{story}
{question}
Choose from the following:
{containers0}, {containers1}.
"""
```

```
Evaluation for StoryLength, QuestionOrder:
baselinePrompt = """\
{story}
{question}
Choose from the following:
{containers0}, {containers1}
{containers2}, {containers3}
{containers4}, {containers5}
{containers6}, {containers7}
{containers8}, {containers9}
{containers10}, {containers11}
{containers12}, {containers13}
{containers14}.
"""
```

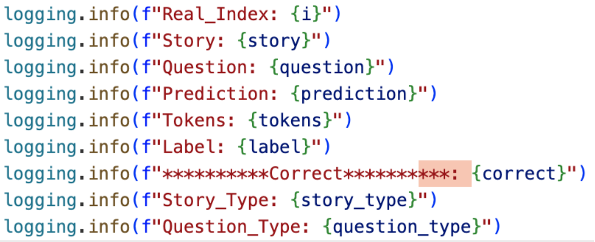
```
Evaluation for ExploreToM:
baselinePrompt = """\
{story}
{question}
Give the answer to this question.
"""
```

```
DeepSeek-v3 Judge Prompt:
prompt = """\
[Question: {question}]

***[Response Answer: {prediction}]***

***[Correct Answer: {label}}]***

Only based on the ***[Correct Answer
]***, judge whether the ***[Response
Answer]*** is correct. Output 'True'or '
False' only.
"""
```

In the process of evaluating the social reasoning capabilities of LLMs, we also established a comprehensive logging system that meticulously records the model's reasoning output for each question, question type, question difficulty, the number of tokens consumed during reasoning, and various other fine-grained metrics.

A.2 Imaginative Game-based Problems

AI Akinator Game: An LLM attempts to determine which object the player is thinking of by asking up to 20 yes-or-no questions. This game demands both deductive and inductive logic capabilities, requires the LLM to think creatively and logically based on the information at hand. We convert the game data into a reasoning problem format, as shown in Figure 5.



```
logging.info(f"Real_Index: {i}")
logging.info(f"Story: {story}")
logging.info(f"Question: {question}")
logging.info(f"Prediction: {prediction}")
logging.info(f"Tokens: {tokens}")
logging.info(f"Label: {label}")
logging.info(f"**********Correct**********: {correct}")
logging.info(f"Story_Type: {story_type}")
logging.info(f"Question_Type: {question_type}")
```

Figure 4: Logging system.

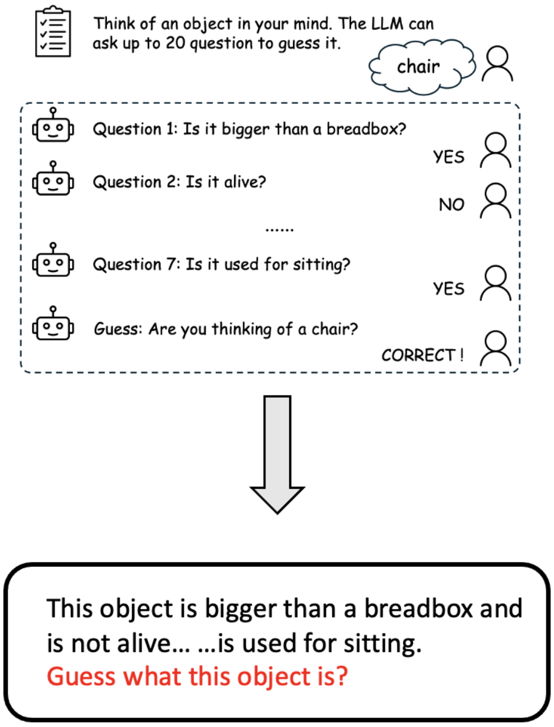

Think of an object in your mind. The LLM can ask up to 20 question to guess it.
chair
Question 1: Is it bigger than a breadbox?
YES
Question 2: Is it alive?
NO
......
Question 7: Is it used for sitting?
YES
Guess: Are you thinking of a chair?
CORRECT !
This object is bigger than a breadbox and is not alive... ...is used for sitting.
Guess what this object is?

Figure 5: Imaginative game-based problems that demand creative logic.

A.3 DeepSeek-R1 Model's Reasoning Trajectory

The DeepSeek-R1 model has an extremely long reasoning trajectory, and the number of reasoning tokens consumed increases significantly as the difficulty of the reasoning problem rises. We have provided the log files in the Data Zip folder.

A.4 Visualization of model size scaling and reasoning tokens scaling

To better visualize the development in social reasoning capabilities during the evolution of model

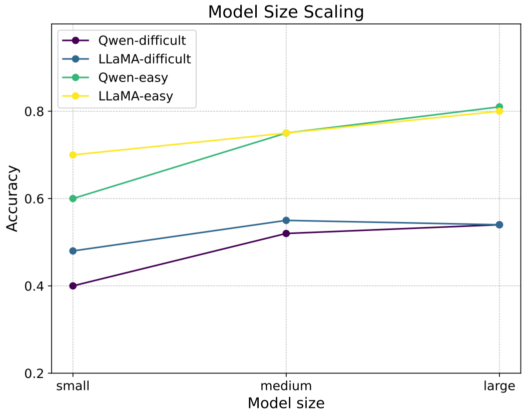


Figure 6: The visualization of model size scaling.

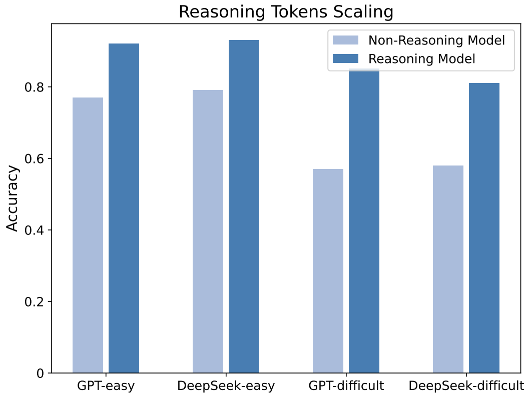


Figure 7: The visualization of reasoning tokens scaling.

size and reasoning tokens, we present the results in Figures 6 and 7.