Pruning for Mindreading

Uncovering Specialized ToM Subnetworks in LLMs via Task-Specific Sparsification

Massimo Stefan
240998
massimo.stefan@studenti.unitn.it

February 6, 2025

Abstract

Large language models (LLMs) have demonstrated remarkable progress across diverse linguistic and cognitive benchmarks, yet questions remain about whether such models truly possess the capacity to reason about mental states—commonly referred to as "Theory of Mind" (ToM)—or merely reproduce surface-level patterns from large-scale training data [5, 14, 16]. In this paper, we investigate whether ToM-related reasoning might be associated with specialized subnetworks or "circuits" in LLMs. By focusing on the possibility that certain subsets of weights are more crucial for tasks invoking mental-state reasoning, we aim to explore how targeted pruning might unveil the presence or absence of domain-specific representations tied to social-cognitive inference.

Our experiments rely on two instruction-tuned models from the Llama-3 family, namely Llama-3.1-8B-Instruct and Llama-3.2-3B-Instruct. We subject them to subtask-specific pruning using Wanda [13]. We pursue moderate (25%) and more aggressive (50%) sparsity levels while calibrating pruning on a small, targeted dataset for each ToM subtask. The pruned variants are then evaluated on ToMBench [1], a comprehensive collection of eight subtasks designed to assess distinct facets of theory-of-mind abilities—from recognizing false beliefs to detecting subtle hints in conversation.

Our results suggest that moderate levels of pruning, such as removing 25% of the weights, often produce minimal disruptions to ToM performance and preserve general language understanding. Unexpectedly, in certain subtasks, more aggressive pruning at 50% not only maintains accuracy but appears to slightly boost performance. This surprising improvement, although not universal across tasks, points to the possibility that specialized "circuits" relevant to certain cognitive processes might be isolated or disinhibited by the pruning [15]. However, we also observe that some tasks, notably the Strange Story Task (SST) [1], exhibit greater sensitivity to pruning, suggesting that distinct parameters may be more crucial for complex social reasoning.

To investigate whether such domain-specific pruning affects other linguistic and cognitive capabilities, we evaluate the pruned models on general benchmarks like Winogrande and ARC. We find that up to 25% sparsity can be achieved with negligible performance loss in standard tasks and a modest increase in perplexity on WikiText. At 50% pruning, we see more pronounced trade-offs in overall language modeling performance, yet certain ToM tasks remain surprisingly robust. These findings collectively illuminate how large models trained on broad data might encode partial, overlapping "circuits" for tasks that appear cognitively specialized, as well as how pruning helps to identify these overlapping but not strictly modular structures. We conclude that analyzing pruning-induced transformations is a revealing way to map the internal foundations of LLMs' social-cognitive reasoning. Our contributions include a set of 16 pruned Llama vari-

ants (eight ToM subtasks at two sparsity levels), along with a comparative analysis of cross-task interactions that can inform future efforts in interpretability, model efficiency, and specialized subnetwork discovery.

1 Introduction

Understanding how people reason about others' beliefs and intentions—commonly grouped under the umbrella of "Theory of Mind" (ToM)—has long been central to fields like developmental psychology, cognitive science, and philosophy of mind. The question of whether large language models (LLMs) possess anything akin to genuine ToM has become increasingly pressing [5, 14, 10]. On the surface, modern LLMs are capable of solving tasks that require apparent social intuition, such as deducing motives from textual vignettes or inferring a protagonist's hidden intentions [11]. Yet whether these successes are driven by emergent patterns of high-level cognition or by simpler, associative pattern matching remains uncertain.

Debates surrounding emergent ToM in LLMs hinge on the fact that purely text-based training could theoretically elicit advanced reasoning capacities, either because the pretraining corpus includes massive amounts of social narrative that indirectly encodes mindreading-like inferences or because large neural networks might spontaneously learn hierarchical abstractions akin to mental-state representations. An opposing perspective posits that LLM performance arises chiefly from lexical and pragmatic correlations: the models excel at reassembling learned patterns in ways that can mimic real reasoning but lack a robust representation of belief, desire, or intention. Multiple studies support both sides. Some find that LLMs succeed on simplified ToM tasks [5, 11], whereas others argue their success dissolves under minor scenario tweaks, subtle rephrasings, or multi-stage contexts where a robust understanding of mental states is crucial [14, 16].

In view of this continuing debate, our research aims to determine whether certain parameters or "circuits" in LLMs specifically underlie ToM capabilities. The notion of domain-specific circuits within neural networks has historical parallels in classical cognitive science, where specialized modules were argued to govern, for instance, language acquisition or face recognition. Contemporary LLMs are not explicitly modular in design but sometimes exhibit localized subnetwork specializations. For instance, prior work on interpretability indicates that pruning or ablating certain attention heads can disrupt particular syntactic or semantic phenomena [2, 15] while leaving other aspects of the model intact. Whether such phenomena extend to social-cognitive abilities such as ToM remains unknown.

Our experimental approach leverages the possibility that targeted pruning can reveal, or partially isolate, domain-relevant parameters. Specifically, we employ "Wanda," a one-shot pruning method designed to remove weights [13] deemed less significant for a particular training or calibration dataset. Wanda's strength lies in its reliance on a small set of calibration samples rather than extensive retraining, making it suitable for large-scale models where iterative methods might be computationally infeasible. By systematically pruning a model on a single ToM subtask and then evaluating it on that subtask as well as others, we can observe whether specialized knowledge is preserved, disrupted, or transferred. This line of inquiry helps clarify whether distinct aspects of social reasoning—like false-belief understanding versus hint interpretation—occupy overlapping or largely disjoint sets of parameters.

We base our analyses on ToMBench [1], which brings together eight different ToM subtasks, each reflecting a specific dimension of mental-state reasoning. These include classic tasks like the False Belief Task (FBT), where an agent's inaccurate mental representation must be tracked, and the Strange Story Task (SST), requiring more nuanced interpretations of social scenarios and potential deceptions. Recent results from ToMBench show that advanced models such as GPT-4 or Llama-2 exhibit performance still falling short of human baselines, especially under coherence tests that reward consistent understanding across multiple queries about the same scenario [1]. Moreover, prior findings suggest that seemingly robust performance can collapse with trivial alterations—like reordering story elements or adding extraneous details—implying that success in these tasks does not necessarily confirm stable internal reasoning akin to human ToM [14].

Pruning as a tool for investigating subnetwork specialization is especially promising, given how widely it is used in practice to reduce model size and inference costs. Techniques like SparseGPT, SlimGPT, and Wanda [3, 8, 13] highlight that large models can remain accurate at moderate sparsity levels, hinting that redundant or partially redundant weights exist. The key twist in our methodology is that we do not simply prune for global model performance or perplexity but instead calibrate the pruning mask on a specific ToM task. We then test how the resulting "ToM-pruned" subnetwork behaves across the entire ToMBench suite and on general tasks such as Winogrande, ARC, and Wikitext perplexity. Our rationale is that analyzing cross-task interactions reveals whether ToM abilities rely on a globally distributed weight pattern or if a smaller fraction of parameters is responsible for mental-state reasoning.

This leads us to three core questions. First, which parts of the LLM—if any—are distinctly important for tasks involving mental-state inference? One possibility is that ToM tasks share a latent subnetwork that is reliably engaged whenever beliefs or intentions are at stake. Another is that these tasks tap into more general language and reasoning pathways scattered throughout the model. Second, does pruning on one ToM subtask degrade performance on the others, or do we see minimal cross-interference? A strong drop in unrelated subtask performance after specialized pruning might indicate that these skills share broad sets of parameters, whereas more selective disruption would suggest subtask-specific knowledge. Third, how does domain-specific pruning affect standard language modeling and general reasoning benchmarks? If a ToM subnetwork is truly isolated, we might expect minimal performance degradation on tasks that do not require complex mindreading. On the other hand, if ToM draws heavily on the same parameters used for general reasoning, pruning might degrade overall language performance.

Our preliminary results reveal nuanced answers that enrich the debate. At 25% sparsity, we observe only marginal performance decline on each subtask, indicating that large portions of the network are not strictly necessary for accurate mental-state inference. Surprisingly, at 50% sparsity, certain tasks—particularly the Strange Story Task—prove more sensitive, while some tasks experience a modest performance boost. We hypothesize this might be due to a "cleaning up" effect: unimportant parameters, possibly containing conflicting or noise-inducing patterns, are removed, making inferences clearer. Not all tasks benefit, however, and the differences among subtasks suggest partial specialization. The Strange Story Task, which covers especially intricate social narratives, seems reliant on a broader set of parameters, so its performance degrades more severely.

An equally important finding is that overall language-model perplexity and standard zero-shot performance remain surprisingly stable up to 25% pruning. This suggests that the region of weight space that handles fundamental language understanding overlaps only partly with the subnetwork needed for certain advanced ToM tasks. However, once we push pruning to 50%, perplexity starts rising more noticeably, and tasks requiring complex, multi-step reasoning, such as the ARC Challenge, begin to suffer. The fact that a purely single-pass pruning strategy like Wanda can preserve so much performance indicates not only the resilience of large transformer architectures but also underscores the potential interpretive significance of subnetwork analysis: even partial removal of parameters can selectively preserve or accentuate mental-state reasoning capabilities in ways that speak to the network's internal distribution of knowledge.

Furthermore, investigating how a subtask-specific pruned model transfers to unrelated tasks, including basic question answering or everyday commonsense reasoning, adds new insight. If pruning for a subtask like False Belief yields minimal performance change in, say, persuasion-story comprehension, that suggests these tasks engage partially distinct parameter sets. By contrast, if performance remains stable on some but not all tasks, we gain a more granular map of how different dimensions of social-cognitive inference might share representational resources.

The broader implications of these findings extend beyond immediate interest in theory of mind. In interpretability research, discovering that removing certain parameters yields stable or even improved performance on specialized tasks might open a path toward more transparent or controllable submodels. For instance, a future line of work could aim to "toggle" a subnetwork specialized for pragmatic implicatures or deception detection in real-world dialogue systems. Efficiency is another domain where subnetwork analysis is relevant, as targeted pruning could reduce model size for deployment scenarios that focus on social reasoning tasks, such as customer service bots or educational tutoring systems, while preserving essential performance.

Ethical considerations also arise. If LLMs can be pruned to selectively preserve or degrade certain social-cognitive functions, developers might inadvertently produce models that are less apt at recognizing manipulative or harmful behavior, or that exhibit biases in how they interpret mental states. A deeper understanding of these subnetworks is crucial for mitigating potential negative consequences and for ensuring that certain beneficial reasoning capacities remain intact.

In the rest of this paper, we detail our methods and findings in steps. Section 2 surveys how researchers have tested ToM in LLMs, the evidence for emergent social-cognitive skills, and how pruning helps isolate crucial parameters in large architectures. Section 3 describes the fundamental concepts behind the Llama-3 model family [4], ToMBench's eight subtasks, and the Wanda pruning technique. Section 4 outlines our experimental design, including model selection, calibration sets, and evaluation metrics for both ToM tasks and general language proficiency. In Section 5, we present quantitative results and interpret how cross-task interference or improvements shed light on potential partial specialization in LLMs. Finally, Section 6 discusses the broader implications of our findings, addresses limitations, and proposes directions for future work that seeks to unravel the complex interplay between domain-specific knowledge and large-scale distributed representations.

2 Related Works

2.1 Theory of Mind Benchmarks for LLMs

Recently, the study of Theory of Mind (ToM) in Large Language Models has become a crucial focus in evaluating how well these models can reason about beliefs, intentions, and other mental states.

- **ToMBench:** [1] introduces a systematic benchmark for theory-of-mind (ToM) in LLMs, containing eight subtasks (False Belief, Hinting, etc.) built from scratch to avoid data contamination. It covers 2,860 multiple-choice questions with bilingual stories. Even strong models (e.g., GPT-4) lag behind human accuracy. Scores drop further under *coherence tests*, suggesting partial or superficial comprehension of mental states.

- **TOMVALLEY:** Another recent dataset [16], featuring 1,100 multi-scenario social contexts with more dynamic and interdependent mental states. This environment requires a model to track how beliefs, intentions, and emotions change across multiple stages, showing that top-tier LLMs still perform below humans.

- **ExploreToM:** [12] programmatically generates adversarial scenarios emphasizing first- and second-order beliefs. GPT-4 and Llama-based models exhibit particularly low success rates when multiple or hidden transfers of objects occur, revealing incomplete ToM.

2.2 Empirical Evidence of Emergent or Fragile ToM

The question of whether LLMs truly exhibit ToM-like cognition or merely approximate it has been widely debated:

- **Pioneering Tests for LLMs:** [5] was the first large-scale attempt to test Theory of Mind abilities in LLMs. Its initial findings caused a stir in the community, as newer models appeared to solve some classic false-belief tasks with surprising accuracy. However, subsequent work with more challenging or non-standard benchmarks revealed that these early successes do not necessarily generalize, leading to substantial performance degradation.
- **Task Robustness:** [14] demonstrates that small logical or phrasing changes can cause performance collapses, hinting at shallow pattern matching rather than robust mental-state modeling.
- **Prompting Benefits:** [11] reports that chain-of-thought prompting substantially boosts certain GPT-based models on ToM tasks, showing LLM outputs can be sensitive to contextual cues.
- **Holistic and Situated Approaches:** [10] and [16] highlight how multi-turn, dynamic contexts reveal deeper failings in LLM belief tracking, underscoring the difference between partial success on short vignettes and consistent "human-like" ToM.
- **Multi-Agent Collaboration:** [7] explores how GPT-4 coordinates with teammates in a text-based bomb-defusal simulation. Models show partial success in attributing beliefs but remain error-prone with second-order reasoning.

2.3 Pruning Approaches in Large Language Models

Pruning has emerged as a major strategy to reduce computational costs in LLMs while attempting to preserve core capabilities:

- **Wanda:** Wanda [13] proposes a simple zero-shot metric that multiplies weight magnitudes by the ℓ_2 norm of input activations. It is highly effective without requiring iterative tuning, matching or outperforming second-order methods;
- **SparseGPT & SlimGPT:** SparseGPT [3] prunes GPT-family models at large scale via a local layer-wise approximation, preserving each layer's input-output relationship. SlimGPT [8] extends structured pruning to entire heads and FFN columns, achieving speedups with near-optimal performance retention;
- **LLM-Pruner & STUN:** LLM-Pruner [9] focuses on structural pruning with minimal data, while STUN [6] blends structured and unstructured pruning in Mixture-of-Experts models, removing entire experts before applying fine-grained weight-level pruning.

Other studies investigate how pruning reveals essential subnetworks or "circuits" for tasks. For example, [15] discusses the effects of removing weights on interpretability metrics. A subnetwork-based approach [2] suggests that relevant linguistic features can live in a sparsified portion of the model.

2.4 The Llama 3 Family of Models

Llama 3 [4] introduces a range of Transformer-based architectures with up to 405B parameters and 128k context windows. Improved data quality and scaling enable advanced performance in language and coding tasks, rivaling GPT-4 on some benchmarks. Instruction-tuned variants (e.g., "-Instruct") incorporate supervised fine-tuning, rejection sampling, and preference optimization. Our experiments leverage mid-sized Llama 3 models (8B, 3B) given resource constraints but are aligned with the general architecture from the broader Llama 3 family.

3 Background

3.1 Language Models

The Llama-3 model family represents state-of-the-art open-source language models featuring high general performance and extensive training. These models undergo conventional large-scale training followed by instruction tuning using Supervised Fine-Tuning (SFT), Rejection Sampling, and Direct Preference Optimization (DPO).

3.2 ToMBench Subtasks

ToMBench [1] is a benchmark covering eight distinct ToM subtasks:

1. Unexpected Outcome Test
2. Scalar Implicature Task
3. Persuasion Story Task
4. False Belief Task
5. Ambiguous Story Task
6. Hinting Test
7. Strange Story Task
8. Faux-Pas Recognition Test

Each subtask includes multiple-choice queries testing comprehension of beliefs, intentions, and implications in realistic scenarios. They vary from basic belief attribution (e.g., False Belief Task) to more complex social interactions (e.g., Persuasion Story, Faux-Pas Recognition).

3.3 Pruning Approach: Wanda

Wanda [13] is a one-shot pruning technique developed specifically for large language models. It operates by multiplying each weight's absolute value by the ℓ_2 norm of its corresponding input activation, collected from a small calibration set. Because Wanda works in a single forward pass, it does not require iterative fine-tuning or second-order calculations. Furthermore, it prunes weights in a balanced manner per output neuron, ensuring a uniform distribution of surviving weights across each layer. This design choice tends to make Wanda minimally perturbative in weight space, often preserving overall performance even at moderate sparsities.

4 Methodology

4.1 Experimental Setup

4.1.1 Model Selection

For this study, we selected two variants of the Llama-3 model family:

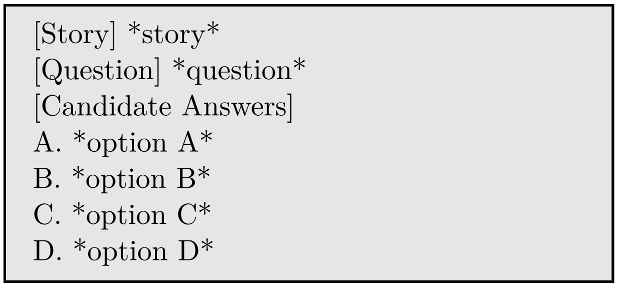
- **Llama-3.1-8B-Instruct**: Our primary base model, featuring high general performance.
- **Llama-3.2-3B-Instruct**: A smaller variant designed for edge applications.

Both models share the same architecture and 128k token context window, making them suitable for direct comparison while representing different points in the compute-performance trade-off spectrum.

4.1.2 Task Configuration

We employ ToMBench subtasks with the following specifications:

- Each task follows a consistent prompt structure with system and user prompts (system prompt shown in Appendix 7).
- Models output only answer indices in [[*]] format without explanations.
- Chain-of-thought prompting was deliberately avoided to focus on core ToM abilities.
- Sample sizes range from 100 to 600 per subtask.
- We use 100 random evaluation samples per subtask.
- When the pruned models on subtask A were evaluated on subtask A, and subtask A had only 100 samples, the evaluation set consisted of only 36 samples.



[Story] *story*
[Question] *question*
[Candidate Answers]
A. *option A*
B. *option B*
C. *option C*
D. *option D*

Figure 1: Template of a ToMBench question. Each test item features a short story, a question, and multiple candidate answers.

4.1.3 Pruning Configuration

For model sparsification, we use Wanda with the following parameters:

- Sparsity levels: 25% and 50%
- 64-sample calibration set per subtask
- Variable-length calibration sequences
- 16 total pruned variants (8 tasks × 2 sparsity levels)

4.2 Evaluation Metrics

We employ a comprehensive evaluation framework incorporating three complementary metric categories to rigorously assess both task-specific and general model capabilities after pruning:

4.2.1 ToMBench Task Performance

Our primary evaluation focuses on multiple-choice accuracy across ToMBench subtasks. We conduct both intra-task analysis (comparing original vs. pruned model performance on the same task) and cross-task analysis to investigate potential interference effects. This allows us to quantify both the preservation of targeted ToM abilities and any unintended impacts on other mental state reasoning capabilities.

4.2.2 General Language Understanding

To assess broader impacts on language understanding, we evaluate zero-shot performance across established benchmarks:

- **Commonsense Reasoning:** HellaSwag and Winogrande assess preservation of general world knowledge and inferential capabilities
- **Question Answering:** ARC (both Easy and Challenge variants) and BoolQ test retention of factual knowledge and binary reasoning

We limit evaluations to 1,000 samples per benchmark to maintain computational efficiency while ensuring statistical significance.

4.2.3 Language Modeling Capability

To quantify changes in foundational language modeling ability, we measure perplexity on the WikiText-2 dataset. This provides:

- A baseline metric for comparing degradation patterns against theoretical predictions from [13]
- Token-by-token analysis revealing any systematic weaknesses introduced by pruning
- Quantitative bounds for expected performance deterioration at different sparsity levels

This multi-faceted evaluation strategy enables us to test our central hypothesis: that task-specific pruning can preserve targeted ToM capabilities while potentially trading off general model performance in a controlled manner.

5 Results & Analysis

5.1 Cross-Task Interactions

Our analysis reveals intricate relationships between pruning effects across different ToM subtasks. The experimental results demonstrate that pruning weights optimized for one specific ToM subtask can have nuanced impacts on performance across other mental state reasoning capabilities. This suggests the existence of shared neural representations supporting multiple aspects of theory of mind reasoning.

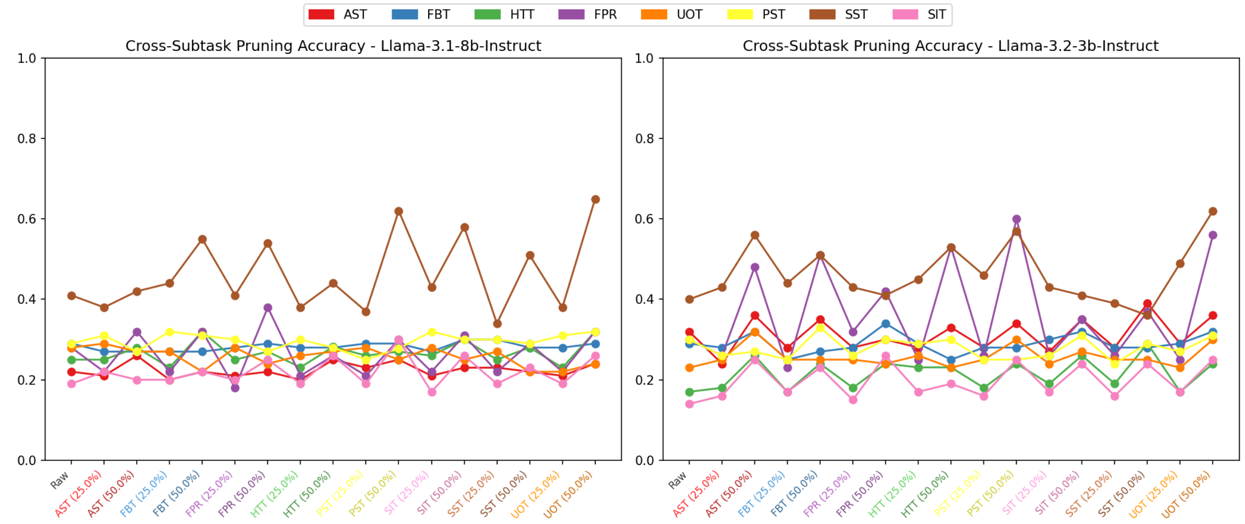


Figure 2: Impact of Subtask-Specific Pruning on ToM Performance. The lines indicate the accuracies on various ToM subtasks (y-axis), while the x-axis represents the raw and pruned models at different sparsity levels.

Notably, the Strange Story Task (SST) exhibited heightened sensitivity to pruning interventions, with accuracy fluctuations significantly larger than other subtasks. This particular vulnerability might indicate that SST relies on more specialized or concentrated neural pathways compared to other ToM capabilities. Despite these variations, most subtasks demonstrated remarkable resilience, maintaining performance within narrow bands - specifically within 0.05 accuracy points for Llama-3.1-8b-Instruct and 0.15 for Llama-3.1-8b-Instruct, even under aggressive 50 percent pruning conditions.

5.2 Analysis of Model Internals and Pruned Structures

Investigation of the internal model architecture post-pruning reveals several key characteristics of Wanda's approach. The pruning mechanism maintains uniform sparsification across all layers, ensuring consistent density reduction throughout the network. This balanced approach appears crucial for preserving model functionality.

Detailed analysis of weight distributions at the layer level shows minimal perturbation in the statistical properties of remaining weights. This observation aligns with Wanda's design philosophy of minimal intervention in weight space. The preservation of weight magnitude distributions likely contributes to the robust maintenance of model performance, particularly in core reasoning capabilities.

5.3 Perplexity and General Language Performance

Our examination of broader language capabilities reveals a nuanced pattern of preservation and degradation under pruning. Zero-shot performance on fundamental reasoning tasks, particularly Winogrande and ARC_Easy, demonstrates notable resilience to pruning. However, the impact of aggressive pruning becomes more pronounced in complex reasoning scenarios.

The ARC Challenge dataset shows particularly steep performance degradation under heavy pruning conditions, suggesting that more complex reasoning paths may be more vulnerable to network sparsification. This observation is further supported by the systematic increase in Wikitext perplexity scores as pruning intensity increases, indicating a gradual erosion of fundamental language modeling capabilities.

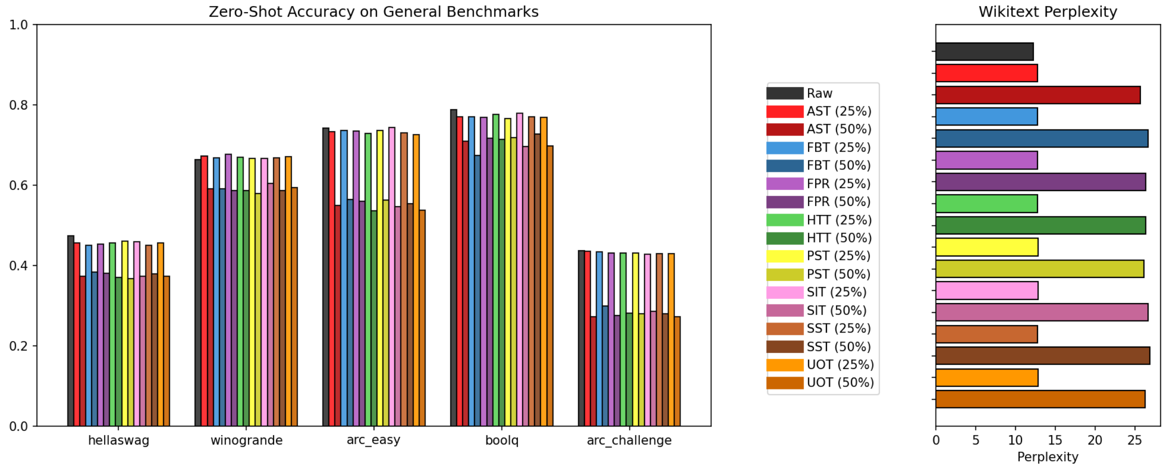


Figure 3: On the left, the zero-shot accuracies on general benchmarks for LLMs; on the right, the Wikitext perplexity on the same models (raw and pruned).

6 Discussion & Conclusion

6.1 Interpreting Task-Specific Pruning for ToM

Our results show that pruning weights based on a single ToM subtask can reveal latent "circuits" responsible for mental-state inference. When a model is calibrated for a specific subtask, it remains surprisingly capable of other ToM tasks, at least at moderate sparsity levels, suggesting there is considerable overlap in the parameters that sustain multiple facets of mindreading. Nonetheless, the observation that certain tasks exhibit greater vulnerability to pruning, especially when we reach or exceed 50% sparsity, indicates that the network also employs subtask-specific weights to address more fine-grained nuances in social reasoning. The Strange Story Task, in particular, appears to require a broad distribution of parameters, perhaps reflecting the multifaceted complexity of sarcasm, deception, and moral inference often encountered in those scenarios [1].

These findings offer a nuanced perspective on the emergent vs. specialized debate [16, 10]. Our evidence does not imply that the entirety of ToM is localized to a discrete, self-contained module. Indeed, each subtask experiences only moderate performance drops even when pruned based on another subtask's calibration data. At the same time, the consistent differences in sensitivity suggest that certain social-cognitive skills require specialized patterns of weights—even if those patterns partially overlap with or draw upon a more general set of linguistic and inferential parameters that the model uses for everyday language tasks.

6.2 Emergent vs. Specialized Social-Cognitive Abilities

One lingering question is whether the partial "circuits" uncovered by subtask-specific pruning reflect emergent general-purpose reasoning or genuine domain specificity. If emergent capabilities were purely an accident of large-scale training, we might expect them to be distributed so widely across the model's parameters that subtask-specific pruning would degrade performance across all tasks [14]. Our observations, however, are more subtle. Cross-task performance remains relatively robust, suggesting that the knowledge relevant to each subtask is at least somewhat distinct or at least partially disentangled from knowledge relevant to others. That independence raises the intriguing possibility that repeated interactions with social-narrative data could cultivate sub-structures that handle mental-state reasoning, even in a model not explicitly designed for that purpose.

At the same time, the phenomenon of moderate pruning leading to stable or even modestly improved accuracy on some tasks [15] might stem from the removal of weights that previously introduced conflicting signals—revealing partial specialization rather than a wholly emergent skill. Thus, LLM-based ToM abilities may be neither purely emergent nor strictly modular. Instead, they seem to lie on a continuum, where certain social-cognitive patterns can be teased out through pruning without entirely divorcing them from the rest of the language or reasoning apparatus.

6.3 Practical Implications for Model Deployment

From a practical standpoint, these findings can inform scenarios where a model might need robust social-cognitive reasoning under computational constraints. If an application depends heavily on, say, persuasion-story comprehension, the developer could prune the model for that subtask to reduce memory and inference costs while still preserving adequate performance in general tasks. Although we see minimal perplexity increases at 25% sparsity, heavier pruning may begin to compromise more advanced reasoning or degrade fundamental skills like language modeling. A real-world system might thus weigh trade-offs: moderate pruning can offer a good balance of efficiency and stable performance, while extreme sparsification might be reserved for highly specialized deployment where only certain tasks matter.

Moreover, the possibility that 50% pruning can improve some subtasks underscores the nontrivial ways in which model capacity and redundancy might interact. In certain corners of the parameter space, reducing extraneous or contradictory weights effectively "sharpens" the relevant subnetwork's focus. This phenomenon could be further tested by applying iterative or partial fine-tuning after pruning, potentially leading to stable improvement without the overhead of training a full-scale model from scratch.

6.3.1 Limitations

Despite these encouraging signs, several limitations should temper our conclusions. First, the two Llama-3 variants we tested (8B and 3B parameters) [4] occupy only the mid-range of current model scales. While our approach is well-grounded, more powerful models may reveal distinct or more intricate patterns of circuit specialization. Second, we used relatively small, carefully curated calibration sets—one per subtask—and a single calibration technique (Wanda). There might be subtler or more effective ways to combine data from multiple subtasks or to iteratively refine pruning masks, potentially exposing deeper subnetwork structures. Furthermore, certain tasks in ToMBench are tested on no more than a few hundred samples [1]. This modest scale constrains the statistical power of our evaluations, even as it enables more fine-grained analysis of cross-task interactions. Larger-scale or more naturalistic ToM datasets, such as multi-turn dialogues or dynamic social contexts [7], may yield richer or more complex insights.

Additionally, it remains unclear how much of the partial "circuitry" discovered here is genuinely about mental-state reasoning vs. more generic narrative comprehension. There is a possibility that each ToM subtask's data distribution, rather than the phenomenon itself, leads the pruning algorithm to preserve relevant patterns. Finally, from a theoretical viewpoint, the capacity of an LLM to approximate ToM does not necessarily entail that it actually "understands" beliefs or intentions in a human-like manner. Our results bear upon how models encode or distribute social reasoning within their parameters, but they do not prove the presence of conscious mental-state attribution.

6.3.2 Future Directions

A few open avenues are especially promising. One involves taking this subnetwork approach into more realistic, multi-turn dialogues. The complexities of real conversation—where knowledge states evolve over time [16] and interact across participants—might reveal specialized weighting that does not appear in single-turn tasks. Another direction is to examine synergy between pruning and lightweight fine-tuning or instruction tuning: if small sub-networks can be "shaped" to perform better on a specific social-cognitive dimension, might it be possible to assemble a suite of specialized sub-networks in a single model, toggling them as needed for different tasks?

We also envision exploring multi-agent collaboration settings, such as text-based team problem solving, in which each agent's knowledge, beliefs, and intentions must be carefully tracked to coordinate success. Pruning might isolate the parameters vital for belief tracking and second-order reasoning in these interactive contexts. Finally, a deeper theoretical question concerns whether certain aspects of social cognition—like reading hidden intentions or identifying sarcasm—rely on partially overlapping sets of parameters. A systematic cross-evaluation approach could prune a model to highlight sarcasm recognition, for example, and then measure how that subnetwork fares on other forms of mindreading.

In short, we foresee a growing research agenda where domain-specific pruning becomes a tool not just for efficiency but also for interpretability, analyzing the extent to which advanced language models rely on specialized parameter subsets for distinct cognitive operations. Theory of mind provides a uniquely revealing lens for these investigations, given how fundamentally it underpins human social interaction and how intensively it has been studied across disciplines. By continuing to integrate state-of-the-art pruning methods, robust benchmarks like ToMBench, and deeper interpretive strategies, we stand to gain meaningful insights into the architectures and training dynamics that yield, or fail to yield, robust social-cognitive reasoning in modern AI systems.

6.3.3 Concluding Remarks

Pruning-oriented approaches can illuminate a hidden topography of knowledge in large language models—a topography that includes partial sub-networks responsible for social-cognitive skills. Our experiments on Llama-3.1-8B-Instruct and Llama-3.2-3B-Instruct show that pruning does not uniformly degrade all mental-state inferences. Instead, moderate pruning can leave mental-state reasoning intact, or even improved, on certain subtasks, while more aggressive pruning uncovers potential overlaps and divergences in how these tasks are encoded. These findings add momentum to the broader notion that large-scale training can yield emergent yet partially specialized skill sets—akin to the historical tension between domain-general and domain-specific theories of human cognition.

Yet for all the promise, caution is essential. The phenomena we observe may depend on model architecture, size, or the particularities of our curated training prompts and evaluation sets [4, 1]. We have also not fully resolved the question of whether these partial sub-networks represent genuine "theory of mind" or simply robust mimicry learned from textual correlations [14]. Nevertheless, the fact that targeted pruning can systematically influence ToM tasks, while leaving general language capacities relatively stable, underscores the utility of examining how distinct tasks are entangled in or divorced from one another within a single large model. We hope that our findings, along with the public release of 16 pruned variants of Llama-3 models, will catalyze further research, both on interpretability strategies and on methodically dissecting the roots of social-cognitive processing in next-generation AI systems.

By embracing pruning as a lens for exploring internal architecture, we can refine our ideas about when, how, and why large language models exhibit or fail to exhibit a kind of rudimentary theory of mind. Although a fully human-like grasp of beliefs and intentions may lie beyond current AI, the partial capacities we observe—combined with the broad resilience of these networks—point the way toward more principled, transparent, and specialized language models. This in turn promises new opportunities for building AI systems that not only generate fluent text but also navigate social reasoning tasks with a clearer, more interpretable foundation.

References

[1] Author(s). Tombench: Benchmarking theory of mind in large language models. *arXiv preprint arXiv:2402.15052*, 2025.

[2] Steven Cao, Victor Sanh, and Alexander M. Rush. Low-complexity probing via finding subnetworks, 2021.

[3] Elias Frantar and Dan Alistarh. Sparsegpt: Massive language models can be accurately pruned in one-shot, 2023.

[4] Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, Aurelien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Roziere, Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak, Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, Christophe Touret, Chunyang Wu, Corinne Wong, Cristian Canton Ferrer, Cyrus Nikolaidis, Damien Allonsius, Daniel Song, Danielle Pintz, Danny Livshits, Danny Wyatt, David Esiobu, Dhruv Choudhary, Dhruv Mahajan, Diego Garcia-Olano, Diego Perino, Dieuwke Hupkes, Egor Lakomkin, Ehab AlBadawy, Elina Lobanova, Emily Dinan, Eric Michael Smith, Filip Radenovic, Francisco Guzmán, Frank Zhang, Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Govind Thattai, Graeme Nail, Gregoire Mialon, Guan Pang, Guillem Cucurell, Hailey Nguyen, Hannah Korevaar, Hu Xu, Hugo Touvron, Iliyan Zarov, Imanol Arrieta Ibarra, Isabel Kloumann, Ishan Misra, Ivan Evtimov, Jack Zhang, Jade Copet, Jaewon Lee, Jan Geffert, Jana Vranes, Jason Park, Jay Mahadeokar, Jeet Shah, Jelmer van der Linde, Jennifer Billock, Jenny Hong, Jenya Lee, Jeremy Fu, Jianfeng Chi, Jianyu Huang, Jiawen Liu, Jie Wang, Jiecao Yu, Joanna Bitton, Joe Spisak, Jongsoo Park, Joseph Rocca, Joshua Johnstun, Joshua Saxe, Junteng Jia, Kalyan Vasuden Alwala, Karthik Prasad, Kartikeya Upasani, Kate Plawiak, Ke Li, Kenneth Heafield, Kevin Stone, Khalid El-Arini, Krithika Iyer, Kshitiz Malik, Kuenley Chiu, Kunal Bhalla, Kushal Lakhotia, Lauren Rantala-Yeary, Laurens van der Maaten, Lawrence Chen, Liang Tan, Liz Jenkins, Louis Martin, Lovish Madaan, Lubo Malo, Lukas Blecher, Lukas Landzaat, Luke de Oliveira, Madeline Muzzi, Mahesh Pasupuleti, Mannat Singh, Manohar Paluri, Marcin Kardas, Maria Tsimpoukelli, Mathew Oldham, Mathieu Rita, Maya Pavlova, Melanie Kambadur, Mike Lewis, Min Si, Mitesh Kumar Singh, Mona Hassan, Naman Goyal, Narjes Torabi, Nikolay Bashlykov, Nikolay Bogoychev, Niladri Chatterji, Ning Zhang, Olivier Duchenne, Onur Çelebi, Patrick Alrassy, Pengchuan Zhang, Pengwei Li, Petar Vasic, Peter Weng, Prajjwal Bhargava, Pratik Dubal, Praveen Krishnan, Punit Singh Koura, Puxin Xu, Qing He, Qingxiao Dong, Ragavan Srinivasan, Raj Ganapathy, Ramon Calderer, Ricardo Silveira Cabral, Robert Stojnic, Roberta Raileanu, Rohan Maheswari, Rohit Girdhar, Rohit Patel, Romain Sauvestre, Ronnie Polidoro, Roshan Sumbaly, Ross Taylor, Ruan Silva, Rui Hou, Rui Wang, Saghar Hosseini, Sahana Chennabasappa, Sanjay Singh, Sean Bell, Seohyun Sonia Kim, Sergey Edunov, Shaoliang Nie, Sharan Narang, Sharath Raparthy, Sheng Shen, Shengye Wan, Shruti Bhosale, Shun Zhang, Simon Vandenhende, Soumya Batra, Spencer Whitman, Sten Sootla, Stephane Collot, Suchin Gururan-

gan, Sydney Borodinsky, Tamar Herman, Tara Fowler, Tarek Sheasha, Thomas Georgiou, Thomas Scialom, Tobias Speckbacher, Todor Mihaylov, Tong Xiao, Ujjwal Karn, Vedanuj Goswami, Vibhor Gupta, Vignesh Ramanathan, Viktor Kerkez, Vincent Gonguet, Virginie Do, Vish Vogeti, Vítor Albiero, Vladan Petrovic, Weiwei Chu, Wenhan Xiong, Wenyin Fu, Whitney Meers, Xavier Martinet, Xiaodong Wang, Xiaofang Wang, Xiaoqing Ellen Tan, Xide Xia, Xinfeng Xie, Xuchao Jia, Xuewei Wang, Yaelle Goldschlag, Yashesh Gaur, Yasmine Babaei, Yi Wen, Yiwen Song, Yuchen Zhang, Yue Li, Yuning Mao, Zacharie Delpierre Coudert, Zheng Yan, Zhengxing Chen, Zoe Papakipos, Aaditya Singh, Aayushi Srivastava, Abha Jain, Adam Kelsey, Adam Shajnfeld, Adithya Gangidi, Adolfo Victoria, Ahuva Goldstand, Ajay Menon, Ajay Sharma, Alex Boesenberg, Alexei Baevski, Allie Feinstein, Amanda Kallet, Amit Sangani, Amos Teo, Anam Yunus, Andrei Lupu, Andres Alvarado, Andrew Caples, Andrew Gu, Andrew Ho, Andrew Poulton, Andrew Ryan, Ankit Ramchandani, Annie Dong, Annie Franco, Anuj Goyal, Aparajita Saraf, Arkabandhu Chowdhury, Ashley Gabriel, Ashwin Bharambe, Assaf Eisenman, Azadeh Yazdan, Beau James, Ben Maurer, Benjamin Leonhardi, Bernie Huang, Beth Loyd, Beto De Paola, Bhargavi Paranjape, Bing Liu, Bo Wu, Boyu Ni, Braden Hancock, Bram Wasti, Brandon Spence, Brani Stojkovic, Brian Gamido, Britt Montalvo, Carl Parker, Carly Burton, Catalina Mejia, Ce Liu, Changhan Wang, Changkyu Kim, Chao Zhou, Chester Hu, Ching-Hsiang Chu, Chris Cai, Chris Tindal, Christoph Feichtenhofer, Cynthia Gao, Damon Civin, Dana Beaty, Daniel Kreymer, Daniel Li, David Adkins, David Xu, Davide Testuggine, Delia David, Devi Parikh, Diana Liskovich, Didem Foss, Dingkang Wang, Duc Le, Dustin Holland, Edward Dowling, Eissa Jamil, Elaine Montgomery, Eleonora Presani, Emily Hahn, Emily Wood, Eric-Tuan Le, Erik Brinkman, Esteban Arcaute, Evan Dunbar, Evan Smothers, Fei Sun, Felix Kreuk, Feng Tian, Filippos Kokkinos, Firat Ozgenel, Francesco Caggioni, Frank Kanayet, Frank Seide, Gabriela Medina Florez, Gabriella Schwarz, Gada Badeer, Georgia Swee, Gil Halpern, Grant Herman, Grigory Sizov, Guangyi, Zhang, Guna Lakshminarayanan, Hakan Inan, Hamid Shojanazeri, Han Zou, Hannah Wang, Hanwen Zha, Haroun Habeeb, Harrison Rudolph, Helen Suk, Henry Aspegren, Hunter Goldman, Hongyuan Zhan, Ibrahim Damlaj, Igor Molybog, Igor Tufanov, Ilias Leontiadis, Irina-Elena Veliche, Itai Gat, Jake Weissman, James Geboski, James Kohli, Janice Lam, Japhet Asher, Jean-Baptiste Gaya, Jeff Marcus, Jeff Tang, Jennifer Chan, Jenny Zhen, Jeremy Reizenstein, Jeremy Teboul, Jessica Zhong, Jian Jin, Jingyi Yang, Joe Cummings, Jon Carvill, Jon Shepard, Jonathan McPhie, Jonathan Torres, Josh Ginsburg, Junjie Wang, Kai Wu, Kam Hou U, Karan Saxena, Kartikay Khandelwal, Katayoun Zand, Kathy Matosich, Kaushik Veeraraghavan, Kelly Michelena, Keqian Li, Kiran Jagadeesh, Kun Huang, Kunal Chawla, Kyle Huang, Lailin Chen, Lakshya Garg, Lavender A, Leandro Silva, Lee Bell, Lei Zhang, Liangpeng Guo, Licheng Yu, Liron Moshkovich, Luca Wehrstedt, Madian Khabsa, Manav Avalani, Manish Bhatt, Martynas Mankus, Matan Hasson, Matthew Lennie, Matthias Reso, Maxim Groshev, Maxim Naumov, Maya Lathi, Meghan Keneally, Miao Liu, Michael L. Seltzer, Michal Valko, Michelle Restrepo, Mihir Patel, Mik Vyatskov, Mikayel Samvelyan, Mike Clark, Mike Macey, Mike Wang, Miquel Jubert Hermoso, Mo Metanat, Mohammad Rastegari, Munish Bansal, Nandhini Santhanam, Natascha Parks, Natasha White, Navyata Bawa, Nayan Singhal, Nick Egebo, Nicolas Usunier, Nikhil Mehta, Nikolay Pavlovich Laptev, Ning Dong, Norman Cheng, Oleg Chernoguz, Olivia Hart, Omkar Salpekar, Ozlem Kalinli, Parkin Kent, Parth Parekh,

Paul Saab, Pavan Balaji, Pedro Rittner, Philip Bontrager, Pierre Roux, Piotr Dollar, Polina Zvyagina, Prashant Ratanchandani, Pritish Yuvraj, Qian Liang, Rachad Alao, Rachel Rodriguez, Rafi Ayub, Raghotham Murthy, Raghu Nayani, Rahul Mitra, Rangaprabhu Parthasarathy, Raymond Li, Rebekkah Hogan, Robin Battey, Rocky Wang, Russ Howes, Ruty Rinott, Sachin Mehta, Sachin Siby, Sai Jayesh Bondu, Samyak Datta, Sara Chugh, Sara Hunt, Sargun Dhillon, Sasha Sidorov, Satadru Pan, Saurabh Mahajan, Saurabh Verma, Seiji Yamamoto, Sharadh Ramaswamy, Shaun Lindsay, Shaun Lindsay, Sheng Feng, Shenghao Lin, Shengxin Cindy Zha, Shishir Patil, Shiva Shankar, Shuqiang Zhang, Shuqiang Zhang, Sinong Wang, Sneha Agarwal, Soji Sajuyigbe, Soumith Chintala, Stephanie Max, Stephen Chen, Steve Kehoe, Steve Satterfield, Sudarshan Govindaprasad, Sumit Gupta, Summer Deng, Sungmin Cho, Sunny Virk, Suraj Subramanian, Sy Choudhury, Sydney Goldman, Tal Remez, Tamar Glaser, Tamara Best, Thilo Koehler, Thomas Robinson, Tianhe Li, Tianjun Zhang, Tim Matthews, Timothy Chou, Tzook Shaked, Varun Vontimitta, Victoria Ajayi, Victoria Montanez, Vijai Mohan, Vinay Satish Kumar, Vishal Mangla, Vlad Ionescu, Vlad Poenaru, Vlad Tiberiu Mihailescu, Vladimir Ivanov, Wei Li, Wenchen Wang, Wenwen Jiang, Wes Bouaziz, Will Constable, Xiaocheng Tang, Xiaojian Wu, Xiaolan Wang, Xilun Wu, Xinbo Gao, Yaniv Kleinman, Yanjun Chen, Ye Hu, Ye Jia, Ye Qi, Yenda Li, Yilin Zhang, Ying Zhang, Yossi Adi, Youngjin Nam, Yu, Wang, Yu Zhao, Yuchen Hao, Yundi Qian, Yunlu Li, Yuzi He, Zach Rait, Zachary DeVito, Zef Rosnbrick, Zhaoduo Wen, Zhenyu Yang, Zhiwei Zhao, and Zhiyu Ma. The llama 3 herd of models, 2024.

[5] Michal Kosinski. Evaluating large language models in theory of mind tasks. *Proceedings of the National Academy of Sciences*, 121(45), October 2024.

[6] Jaeseong Lee, seung-won hwang, Aurick Qiao, Daniel F Campos, Zhewei Yao, and Yuxiong He. Stun: Structured-then-unstructured pruning for scalable moe pruning, 2024.

[7] Huao Li, Yu Chong, Simon Stepputtis, Joseph Campbell, Dana Hughes, Charles Lewis, and Katia Sycara. Theory of mind for multi-agent collaboration via large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 2023.

[8] Gui Ling, Ziyang Wang, Yuliang Yan, and Qingwen Liu. Slimgpt: Layer-wise structured pruning for large language models, 2024.

[9] Xinyin Ma, Gongfan Fang, and Xinchao Wang. Llm-pruner: On the structural pruning of large language models, 2023.

[10] Ziqiao Ma, Jacob Sansom, Run Peng, and Joyce Chai. Towards a holistic landscape of situated theory of mind in large language models, 2024.

[11] Shima Rahimi Moghaddam and Christopher J. Honey. Boosting theory-of-mind performance in large language models via prompting, 2023.

[12] Melanie Sclar, Jane Dwivedi-Yu, Maryam Fazel-Zarandi, Yulia Tsvetkov, Yonatan Bisk, Yejin Choi, and Asli Celikyilmaz. Exploretom: program-guided adversarial data generation for theory of mind reasoning. 2024. under review.

[13] Mingjie Sun, Zhuang Liu, Anna Bair, and J. Zico Kolter. A simple and effective pruning approach for large language models, 2024.

[14] Tomer Ullman. Large language models fail on trivial alterations to theory-of-mind tasks, 2023.

[15] Jonathan von Rad and Florian Seuffert. Investigating the effect of network pruning on performance and interpretability, 2025.

[16] Yang Xiao, Jiashuo WANG, 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, 2024.

7 System Prompt

Below is a multiple-choice question with a story and serveral answer options. Based on the content of the story and the given question, please infer the most likely answer and output the answer index. Note:

(1) Please only output the most likely answer index in the format: [[Answer Index]], for example, if the most likely answer option is 'A. Handbag', then output '[[A]]';

(2) You must choose one of the given answer options 'A, B, C, D' as the most likely answer, regardless of whether the story provides enough information. If you think there is not enough information in the story to choose an answer, please randomly output one of "[[A]]", "[[B]]", "[[C]]", or "[[D]]";

(3) Please only output the most likely answer index based on the given information, and do not output any other content.

8 Examples ToM Subtasks

8.1 Example data of the Unexpected Outcome Test (UOT)

Story: Xiao Ming receives a bicycle on his birthday.

Question-1: What is Xiao Ming's emotion?
(A) Embarrassed
(B) Happy
(C) Disappointed
(D) Regretful

Question-2: He should be very happy, but he is very disappointed, why?
(A) Xiao Ming worries that riding a bicycle affects his studies.
(B) Xiao Ming fears that riding a bicycle to school makes his classmates laugh at him.
(C) Xiao Ming thinks the color of the bicycle does not match his clothes.
(D) Xiao Ming hopes for a computer as a gift, not a bicycle.

8.2 Example data of the Scalar Implicature Task (SIT)

Story: Almost every letter to Laura Company contains a check. Today, Laura receives 5 letters. Laura tells you on the phone “I look at 3 out of 5 letters. There are checks in 2 of the letters.”

Question-1: Before Laura calls you, how many of these 5 letters do you think contain checks?
(A) 0
(B) 1
(C) 2
(D) 4

Question-2: After Laura calls you, how many of these 5 letters do you think contain checks?
(A) 0
(B) 1
(C) 2
(D) 4

8.3 Example data of the Persuasion Story Task (PST)

Story: Xiao Hong is a 6-year-old child. Today is Saturday. Mom and Dad have free time today, they do not know what they can do together. Maybe go for ice cream? Xiao Hong really, really wants to go to the amusement park today. However, Dad thinks the amusement park has a lot of noise. He says: “Xiao Hong, this is not a good idea. I think there is a lot of noise in the amusement park.”

Question: How does Xiao Hong persuade her dad?
(A) Xiao Hong can look up some information, proving that the current amusement parks take many measures to reduce noise, such as setting up soundproof walls, using quieter equipment, etc.
(B) Xiao Hong can tell her dad that she has not been to the amusement park for a long time, this is a very special wish for her, she really wants to go.
(C) If Dad really does not want to go to the amusement park, Xiao Hong can suggest going to other places, such as parks or zoos, so everyone is happy.
(D) Xiao Hong can suggest: “How about we go in the early morning or on weekdays? I hear that there are fewer people at that time, and it is quieter. We can avoid peak hours, and get a quieter experience.”

8.4 Example data of the False Belief Task (FBT)

Story: Li Lei and Han Meimei are wandering in the living room, they see the cabinet, box and handbag, they find a sweatshirt in the cabinet, Han Meimei leaves the living room, Li Lei moves the sweatshirt to the handbag.

Question-1: Where is the sweatshirt now?
(A) Briefcase
(B) Box
(C) Cabinet
(D) Handbag

Question-2: After Han Meimei returns to the living room, where does Li Lei think Han Meimei looks for the sweatshirt?
(A) Box
(B) Wardrobe
(C) Handbag
(D) Cabinet

8.5 Example data of the Ambiguous Story Task (AST)

Story: Dehua and Sanming are company employees, they are competing for a promotion opportunity. Today is Sanming's birthday, he enjoys a small birthday cake alone. Lingling is the company supervisor, she is having a private conversation with Dehua in the pantry. Sanming sees from a distance that Lingling quietly gives Dehua a document, and with a smile, gently pats Dehua's back, then returns to her office. Dehua quickly checks the document, shows a smile, and carefully puts the document into his file bag.

Question-1: Why does Lingling give Dehua that document?
(A) Lingling returns Dehua's personal document left in the conference room.
(B) Lingling gives Dehua a daily notice unrelated to promotion.
(C) Lingling and Dehua are planning a surprise for Sanming.
(D) Lingling gives Dehua extra information.

Question-2: What do you think Sanming thinks?
(A) Sanming thinks everything is normal because that is Dehua's document.
(B) Sanming feels angry because he is isolated.
(C) Sanming feels very happy because his cake tastes good.
(D) Sanming reevaluates his promotion opportunity.

8.6 Example data of the Hinting Test (HT)

Story: Cao Sheng and Wang Hong are a couple. One day, Wang Hong makes dinner, Cao Sheng takes a bite and says, “Does our home not buy salt for a long time?”

Question: What does Cao Sheng really want to say when he says this?
(A) Cao Sheng is asking if the salt in the house is used up.
(B) Cao Sheng is expressing his increased need for salt.
(C) Cao Sheng implies that the dinner Wang Hong makes does not have enough salt.
(D) Cao Sheng is reminding them that they need to buy more ingredients.

8.7 Example data of the Strange Story Task (SST)

Story: One day, Lily plays at home and accidentally knocks over and breaks her mother’s favorite crystal vase. Ohno, her mother definitely gets angry when she finds out! Therefore, when Lily’s mother comes home, sees the broken vase and asks Lily what happens, Lily says,“The dog knocks it over, it is not my fault!”

Question-1: Is what Lily says true?
(A) Yen
(B) No

Question-2: Why does Lily say this?
(A)Lily sees the dog knock over the vase and wants to protect the dog from her mother’s blame.
(B)Lilyj okes to protect herself from her mother’s blame.
(C)Lily lies to protect the dog from her mother’s blame.
(D)Lily lies to protect herself from her mother’s blame.

8.8 Example data of the Faux-Pas Recognition Test (FRT)

Story: Last Friday,all classmates in the fifth grade of elementary school participate in a Chinese composition competition. Xiao Ming really wants to win first place in this competition. The competition results come out, Xiao Hong wins the competition,and Xiao Ming happens to take a leave of absence and does not go to school that day. When going to school on Monday, Xiao Hong is the first to see Xiao Ming and says:“Xiao Ming,I read your composition and it is not bad, just a bit regrettable.”Xiao Ming is very surprised when he hears this and says: ”What do you mean, can you be more clear? ”Xiao Hong look sat Xiao Ming and says:“Oh, it is nothing, let’s go to class.”

Question-1: Does anyone say something inappropriate in this story?
(A)Someone in the story says something inappropriate.
(B)No one in the story says something inappropriate.

Question-2: Who wins the composition competition?
(A)Xiao Ming
(B)Xiao Long
(C)Xiao Hong
(D)The story does not mention