REPRESENTATION-ALIGNMENT IN THEORY-OF-MIND TASKS ACROSS LANGUAGE MODELS/AGENTS

Rohini Elora Das
New York University
rohini.elora.das@nyu.edu

Krissh Bhargava
Georgia Institute of Technology
krissh.bhargava@gmail.com

Krishna Shinde
Enkefalos Technologies
krishna.shinde@enkefalos.com

Rajarshi Das
MQube Cognition
rajarshi.das@mqube.ai

ABSTRACT

As AI systems increasingly tackle complex reasoning tasks, understanding how they internally structure mental states is crucial. While prior research has explored representational alignment in vision models, its role in higher-order cognition remains under-examined, particularly in Theory of Mind (ToM) tasks. This study evaluates how AI models encode and compare mental states in ToM tasks, focusing on tasks such as False Belief, Irony, and Faux Pas reasoning. Using a triplet-based similarity framework, we assess whether structured reasoning models (e.g., DeepSeek R1) exhibit better alignment than token-based models like LLaMA. While AI models correctly answer individual ToM queries, they fail to recognize broader conceptual structures, clustering stories by surface-level textual similarity rather than belief-based organization. This misalignment persists across 0th, 1st, and 2nd-order ToM reasoning, highlighting a fundamental gap between human and AI cognition. Moreover, explicit reasoning mechanisms in DeepSeek do not reliably improve alignment, as models struggle to capture hierarchical ToM structures. To further probe this gap, we propose extending representational analysis to temporally evolving, multi-agent belief systems—capturing how beliefs about beliefs shift across time and interaction. Our findings suggest that achieving deeper AI alignment requires moving beyond task accuracy toward developing structured, human-like mental representations. Using triplet-based alignment metrics, we propose a novel approach to quantify AI cognition and guide future improvements in reasoning, interpretability, and social alignment. Additionally, we propose this representational framework as a potential foundation for a noninvasive, scalable cognitive monitoring tool for early-stage dementia or Alzheimer's, analogous to fMRI-based biomarkers but deployable through everyday interactions on mobile platforms.

1 INTRODUCTION

As AI systems become more integrated into human-centric applications, understanding the extent to which these models align with human cognition is increasingly critical. While much of AI alignment research has centered on value alignment, representational alignment—how AI models internally represent the world in comparison to humans—is equally essential. When AI systems construct representations that resemble human cognitive structures, they may not only generalize more effectively but also exhibit improved reasoning and interpretability. Given that human representations of the world are typically structured to facilitate efficient learning and inference, representational alignment can serve as a valuable inductive bias, enabling AI models to learn robustly from limited data.

This paper explores the role of representational alignment in cognitive reasoning tasks, focusing on Theory of Mind (ToM) settings such as Irony and False Belief reasoning. ToM refers to the ability to infer and track one's own and others' mental states, a fundamental aspect of human cognition that

underpins social reasoning and communication. Previous research has used ToM-based assessments to compare human and AI performance in complex reasoning tasks (Wang et al., 2024; Strachan et al., 2024). We extend this line of inquiry to evaluate how AI models construct and compare mental representations across different narratives. By systematically analyzing representational similarity, we aim to determine whether alignment with human cognitive structures enhances AI's ability to infer beliefs, intentions, and perspectives in diverse reasoning contexts.

To investigate this, we develop a structured experimental framework for assessing how AI models with and without explicit reasoning mechanisms encode and compare mental states. Building on prior work in representational alignment (Sucholutsky & Griffiths, 2023), we analyze whether models like DeepSeek R1, which incorporate structured reasoning, differ from token-based models like LLaMA in their approach to ToM tasks such as Irony and False Belief. This framework offers a structured approach to examining how an agent may organize representations in cognitive tasks.

In future extensions, we argue that representational alignment should not only be evaluated statically but also temporally, examining how an agent's internal map of nested beliefs evolves over sustained interactions, long-form dialogues, or episodes of social exchange. Such dynamic modeling better captures the recursive, time-sensitive nature of human mental modeling and holds promise for aligning AI agents with human theories of interaction and memory-based social reasoning. Moreover, in addition to helping improve model performance and interpretability, such representational alignment frameworks may also contribute to translational applications in cognitive health. Previous work in neuroscience has used fMRI to track Theory of Mind (ToM) reasoning as a transdiagnostic marker of cognitive decline (e.g. in Alzheimer's disease) (De Lucena et al., 2020; Zegarra-Valdivia et al., 2023; Tripathi et al., 2025). Our alignment-based framework opens the possibility of capturing similar patterns through language-based reasoning and changes in social-cognitive structures over time, enabling novel, scalable tools for early detection and longitudinal assessment of neurodegenerative disease.

2 MOTIVATIONS

Understanding how AI models align with human representations is crucial to improving reasoning and generalization in cognitive tasks. Inspired by prior work in representational alignment in vision models, we extend this investigation to ToM reasoning, where AI or human agents infer and compare mental states across different stories. Representational alignment, while having been explored for vision models (Constantinescu et al., 2016; Das & Das, 2024), remains underexamined in higher-level reasoning tasks such as ToM, Faux Pas, Irony, and False Belief. Understanding how AI agents construct these representations is critical for enhancing both their reasoning and the intuitive manner in which humans interact with them.

To address this gap, we adopt a structured pipeline (Figure 1) comprising:

- **Shared Data** – Present a collection of stories (e.g., involving Alice and Bob) that illustrate multiple mental state scenarios (ToM, Faux Pas, Irony, False Belief).
- **Representations** – Have human observers and AI models (e.g., DeepSeek R1, with explicit reasoning, and LLaMA, without it) encode the stories.
- **Pairwise Similarity** – Compare the resulting encodings to assess how closely they align with human judgments, using a triplet-based framework (Jamieson & Nowak, 2011) to determine whether Story A is more similar to Story B than to Story C.

We have two main goals with this study: (1) determine whether models with explicit reasoning capabilities demonstrate superior alignment and whether substantial misalignment or high alignment might outperform moderate alignment in challenging reasoning tasks, and (2) examine how models construct representations of others' thoughts and whether their inferred mental model similarities resemble human judgments. To operationalize this, we draw inspiration from Sucholutsky & Griffiths (2023), who propose a triplet-based approach to constructing representation spaces. Specifically, we aim to extract a triplet-based representation space in 2D from similarity triplets of the form: "Is Story A closer to Story B than to Story C?"

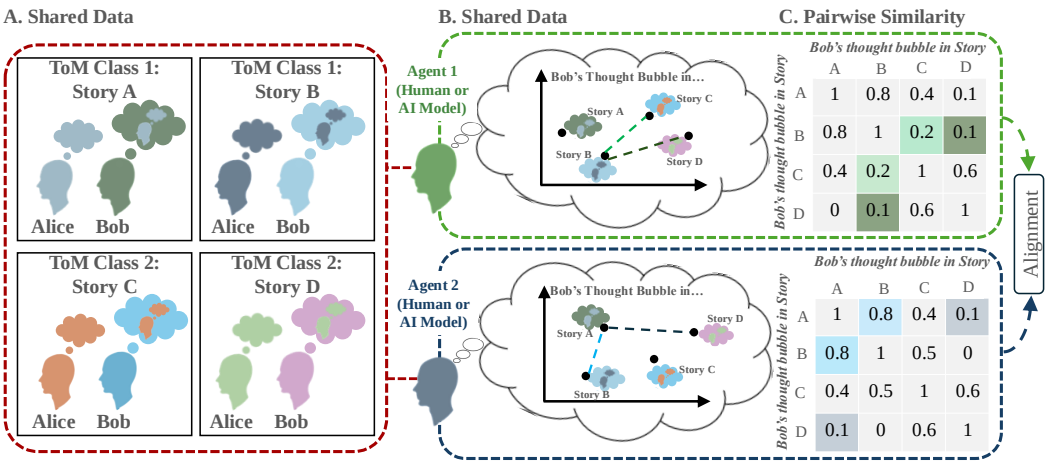


Figure 1: **Schematic of representational alignment between two agents, Agent 1 and Agent 2 (Human or AI Model). A:** Shared ToM stories (x) are shown to both agents (stories are drawn from an existing dataset, e.g., (Strachan et al., 2024)). **B:** Both agents form internal representations ($f_A(x)$ and $f_B(x)$) at different levels of belief (0th, 1st, and 2nd order) of the ToM stories they are given. **C:** Agents are asked to produce pairwise similarity matrices corresponding to their internal representations. The similarity judgments can then be compared to measure alignment between the agents. This figure is adapted from Sucholutsky & Griffiths (2023), which examined representational alignment in vision models, extending the framework to higher-level cognitive tasks.

2.1 Preliminary Investigation

Before focusing on smaller-scale triplet comparisons, we conducted a pilot study using the rich emotional dynamics of *Inside Out 2* (Disney). We prompted several language models—OpenAI's GPT_o1, GPT_4o, and GPT_4—with questions progressing from first-order (e.g., "What does *Anxiety* [1st character] think of *Joy* [2nd character]?") to fourth-order ("What does *Joy* think of how *Anxiety* thinks *Sadness* [3rd character] thinks of *Embarrassment* [4th character]?") in order to gauge how reasoning models, current token-based models, and legacy models respectively deal with higher-order thinking.

GPT_o1 demonstrated a coherent breakdown of higher-order prompts, examining each emotional perspective step by step. For example, when asked about *Joy's* multi-layered view of *Anxiety*, *Sadness*, and *Embarrassment*, GPT_o1 referenced a chain of thought that outlined:

> "Joy → (how *Anxiety* thinks of) Sadness → (how *Sadness* thinks of) *Embarrassment*"

The model then broke the problem down into three sub-components:

1. "What does *Anxiety* generally think of *Sadness*?"
2. "How does *Sadness*, in turn, treat or view *Embarrassment*?"
3. "How does *Joy* feel about all of that happening?"

This chain-of-thought approach highlighted specific relationships among the characters, illustrating an organized, cohesive mental model (OpenAI ChatGPT_o1).

OpenAI ChatGPT_4o and OpenAI ChatGPT_4 on the other hand, offered incomplete or tangential responses, often missing critical links in the nested reasoning. While they identified some relevant character dynamics—for instance, noting that "Joy is always aiming to ensure Riley's happiness and positivity, managing other emotions to maintain balance"—they generally failed to connect Sadness

and Embarrassment with Anxiety's perspective in a cohesive manner. These gaps suggest weaker internal alignment with the intricate multi-step reasoning required by the fourth-order query.

Such discrepancies highlight the value of representational alignment: the more a model's internal reasoning structure mirrors human conceptualizations of nested mental states, the more effectively it can answer complex prompts. Conversely, models that lack explicit reasoning mechanisms often struggle to integrate multiple layers of perspective-taking.

2.2 Moving Forward: Quantifying Alignment

Given our preliminary insights, we now aim to quantify how AI models encode these nuanced mental states. Drawing on Jamieson & Nowak (2011), we use a triplet-based approach to learn a structured representation space, asking whether Story A is more similar to Story B than to Story C. By applying this framework to stories involving ToM, Faux Pas, Irony, and False Belief, we can compare how different architectures—particularly those with explicit reasoning (e.g., DeepSeek R1)—achieve or fail to achieve representational alignment.

Ultimately, this investigation seeks to illuminate the following:

- When and why AI systems succeed in aligning with human judgments, particularly for complex social scenarios.
- How explicitly structured reasoning modules might bolster or bias a model's capacity to form accurate "mental maps."
- How different LLMs visually represent mental states, particularly focusing on differences between reasoning- and token-based models, and how we can systematically model the processes by which they construct mental maps.
- A framework to uncover how an AI agent models A's belief about B's belief, and quantify the evolution of those beliefs in a coherent fashion over long sequences of interactions and time.

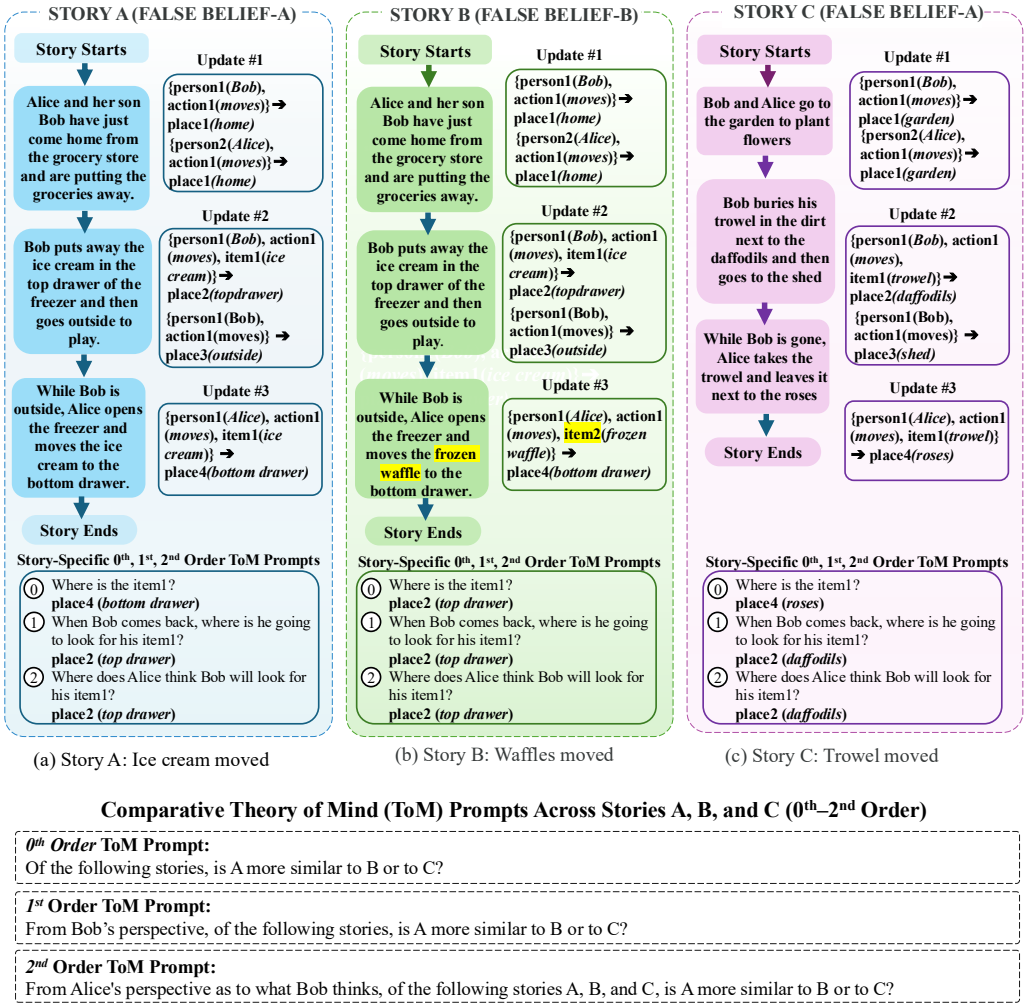
Through this work, we aspire to enhance the design of AI systems that not only provide contextually relevant answers but also reason in a manner that users find more intuitive—advancing both the science of cognitive modeling and practical human–AI interaction.

3 Methods

We sourced ToM questions from Strachan et al. (2024), which assessed ToM reasoning in both humans and large language models. The dataset, initially in spreadsheet format, was adapted into an (Alice, Bob) structure while maintaining original roles and question categories. Specifically, we extracted and reformatted False Belief (A, B), Irony (A, B), and Faux Pas (A) tasks to ensure consistency in narrative presentation across models and human participants.

For False Belief and Irony tasks, each pair of stories (A, B) is nearly identical, differing by only one or two key words that alter the correct response. As illustrated in Figure 2, in Story A, the misplaced object (e.g., ice cream) remains the focal point of the question, whereas in Story B, a minor wording change (e.g., frozen waffles) results in different 0th, 1st, and 2nd order beliefs. The key objective is to determine whether an AI agent can detect these subtle distinctions and correctly judge that Story A is more similar to Story C than to Story B, as both A and C involve reasoning about the same object undergoing a state change.

In Figure 3, the left panel illustrates different classes of ToM stories/scenarios/beliefs, including False Belief A, False Belief B, Irony A, Irony B, and Faux Pas, with variables representing specific stories within each class Strachan et al. (2024). Each story is encoded into an embedding by the agent. Notably, some stories from different classes are textually similar (indicated by connections), leading to highly similar embeddings. The middle panel shows the triplet-based reasoning task, where agents (DeepSeek R1, LLaMA, and humans) are prompted with three key questions: (i) Is Story A more similar to B or C? (0th-order reasoning), (ii) From Bob's perspective, is Story A more similar to B or C? (1st-order reasoning), and (iii) From Alice's perspective on Bob's reasoning, is Story A more similar to B or C? (2nd-order reasoning). Here, the correct answer is the story that



(a) Story A: Ice cream moved (b) Story B: Waffles moved (c) Story C: Trowel moved

Comparative Theory of Mind (ToM) Prompts Across Stories A, B, and C (0th–2nd Order)

***0th Order* ToM Prompt:**
Of the following stories, is A more similar to B or to C?

***1st* Order ToM Prompt:**
From Bob's perspective, of the following stories, is A more similar to B or to C?

***2nd* Order ToM Prompt:**
From Alice's perspective as to what Bob thinks, of the following stories A, B, and C, is A more similar to B or to C?

Figure 2: Three False Belief scenarios: each differs subtly in object manipulation, impacting the inferred mental states in ToM reasoning tasks. Example of False Belief tasks where stories (A, B) differ by a single key word, altering the correct response. The task assesses whether agents can detect these subtle distinctions and correctly infer that Story A is more similar to Story C than to Story B, as both A and C involve reasoning about the same object undergoing a state change.

belongs to the same class as A, rather than the story that initially had a similar embedding. High textual similarity alone does not imply that the key object in the story underwent the same state change. The right panel shows the representational alignment metrics adopted from Sucholutsky & Griffiths (2023) using similarity triplets which gives us the relative magnitude of the pairwise distances. Ideally, after remapping, aligning semantically similar narratives (stories a part of the same class) become more coherently grouped in the representational space by their ToM class affiliation.

3.1 From Triplets to Distance Matrices via Triplet Loss

Representational Alignment: Representational alignment quantifies the degree to which two (or more) agents—often a human and a machine—encode shared structure in their internal representations. Early methods for recovering human similarity spaces date back to Shepard (1980), who advocated using behavioral data (e.g., pairwise or triplet comparisons) to reveal how individuals map stimuli into mental spaces. Multidimensional scaling (MDS) then embeds these similarity relationships into a low-dimensional manifold that respects the observed ordinal constraints.

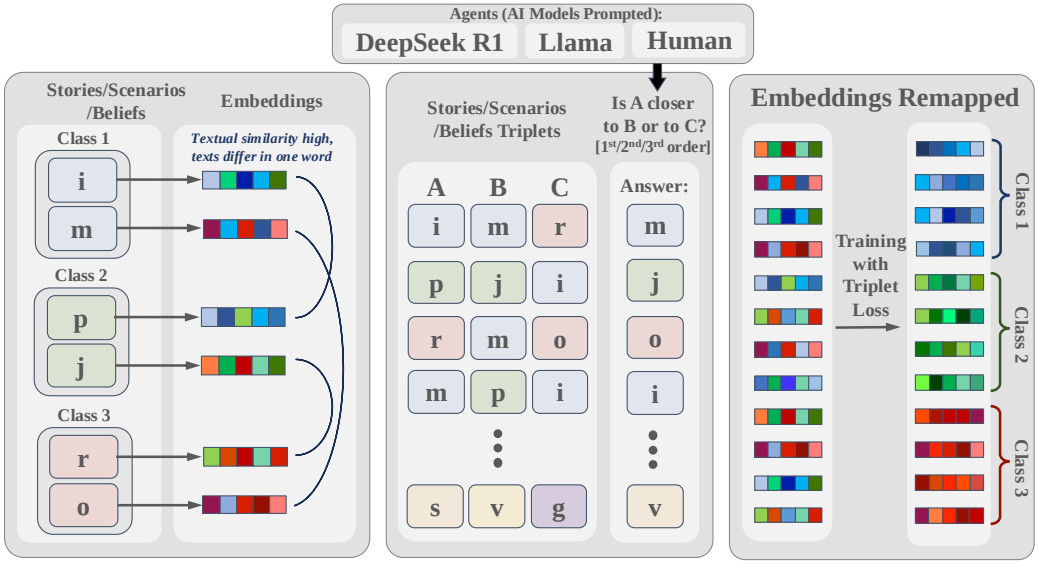


Figure 3: **Schematic representation of the triplet-based reasoning framework.** The left panel shows ToM story classes and their embeddings, the middle panel presents the analogy task across different reasoning orders, and the right panel depicts remapped embeddings, aligning narratives by ToM class affiliation.

Triplet Queries: To efficiently collect similarity judgments, Jamieson and Nowak Jamieson & Nowak (2011) introduced a framework that asks queries of the form: "Is item i more similar (closer) to item j or item k?". They further derived theoretical bounds on how many such queries suffice to recover the entire similarity structure. While perfect recovery may require many queries, approximate methods can still align representations meaningfully. For instance, Peterson et al. Peterson et al. (2018) leveraged pre-trained computer vision models to approximate human perceptual similarities over images, indicating that certain learned embeddings can already reflect important facets of human judgment.

Margin-Based Triplet Criterion: Following this literature and adapting the approach from the Sucholutsky & Griffiths (2023) work, we employ a triplet loss to embed our ToM stories. Specifically, each triplet consists of anchor ($\mathbf{a}$), positive ($\mathbf{p}$), and negative ($\mathbf{n}$) examples, reflecting the judgment that $\mathbf{a}$ is more similar to $\mathbf{p}$ than to $\mathbf{n}$. Mathematically, given embeddings $\mathbf{z}_a = f(\mathbf{a})$, $\mathbf{z}_p = f(\mathbf{p})$, and $\mathbf{z}_n = f(\mathbf{n})$, we define:

$$\mathcal{L}_{\text{triplet}} = \sum_{\text{triplets}} \max\Big(0,\ \alpha + d(\mathbf{z}_a, \mathbf{z}_p) - d(\mathbf{z}_a, \mathbf{z}_n)\Big), \tag{1}$$

where $d(\cdot, \cdot)$ denotes the distance between embeddings (e.g., Euclidean), and α is a small margin. By enforcing $d(\mathbf{z}_a, \mathbf{z}_p) + \alpha \leq d(\mathbf{z}_a, \mathbf{z}_n)$, the anchor–positive pairs are pulled closer than anchor–negative pairs in the learned space.

Distance Matrix Construction. Once the embeddings satisfy this triplet-based criterion, we compute the final $N \times N$ distance matrix D by measuring $d(\mathbf{z}_i, \mathbf{z}_j)$ for all story pairs (i, j). This matrix captures the global geometry of inter-story relationships, reflecting both the local constraints introduced by each triplet and the broader manifold structure. As discussed in prior work (Shepard, 1980; Jamieson & Nowak, 2011; Peterson et al., 2018), such a matrix can be compared across agents—e.g., LLaMA, DeepSeek R1, and human participants—to quantify representational alignment in ToM and related tasks.

4 RESULTS

Figure 4 illustrates the representational alignment between three agents—LLaMA 3.2 (Figure 4c), DeepSeek (DeepSeek-R1-Distill-Llama-8B GGUF:Q8_0) (Figure 4d), and humans (undergraduate students) (Figure 4e)—on ToM stories. The stories are drawn from two distinct classes: False Belief A (Stories 1–8) and False Belief B (Stories 9–17).

Every story in False Belief A has a corresponding counterpart in False Belief B, differing only by one or two words. Since we use DistilBERT to generate initial embeddings for all stories, this high textual similarity is reflected in the smaller diagonals of the distance matrix, where stories with minimal wording differences exhibit strong similarity. However, because each story involves distinct settings and objects, overall similarity between different stories remains low, as shown in (a) of Figure 4.

Based on the same 50 similarity triplets, the representational alignment matrix in Figure 4 reveals key differences in how humans and AI models recover class distinctions in ToM tasks, specifically False Belief A and False Belief B. This alignment is evident in two primary ways:

1. The smaller side diagonals, which previously reflected textual similarity between paired stories, largely disappear in the human distance matrix.
2. Stories within each class (False Belief A and False Belief B) exhibit a higher degree of internal similarity, as observed in the clustering of stories in the top-right and bottom-left quadrants of the matrix.

Moreover, this structured representational alignment persists across all levels of ToM reasoning—0th order, 1st order, and 2nd order—indicating that human participants consistently recognize and maintain these distinctions. This is illustrated in the two "clusters" seen in the bottom right and top left quadrants of the 2nd order human agent matrix (e) in Figure 4.

In contrast, in (d) of Figure 4 the DeepSeek AI model demonstrates only partial alignment as shown by the lack of clustering in the bottom right and top left quadrants. While some reduction in similarity between paired stories and the disappearance of side diagonals is observed, the model largely fails to recover the class affiliations of False Belief A and False Belief B. This misalignment is consistent across 0th, 1st, and 2nd order ToM reasoning, suggesting that while the model may correctly answer individual ToM questions, it struggles to identify broader structural patterns linking stories within the same class.

Notably, despite DeepSeek's explicit reasoning capabilities—where it partitions tasks into sequential subtasks—this approach does not lead to a meaningful improvement in recognizing hierarchical ToM reasoning structures. Similarly, the LLaMA model in (c) of Figure 4, an autoregressive transformer, exhibits comparable difficulties in identifying class structure, reinforcing the limitations of current language models in recovering representational alignment in complex reasoning tasks.

These findings highlight a fundamental gap between human and AI reasoning: while AI models may achieve high task performance on individual ToM queries, they struggle to capture the underlying conceptual organization necessary for coherent mental state inference. This suggests that improving representational alignment, rather than just task accuracy, is critical for advancing AI's ability to perform structured reasoning in social and cognitive domains.

5 CONCLUSION AND FUTURE DIRECTIONS

In this work, we explore representational alignment in the context of Theory of Mind (ToM) by examining how AI systems, particularly those with and without reasoning mechanisms, internally structure their state representations. While prior investigations in representational alignment have often centered on visual domains, our focus on mental states and varying orders of belief (e.g., 0th-, 1st-, 2nd-, and higher-order ToM) reveals new insights into how large language models can capture the nuanced, layered nature of human social cognition. Our pilot study highlighted discrepancies among language models in their ability to integrate nested perspectives, underscoring the importance of structured reasoning modules to achieve ToM-aligned representations.

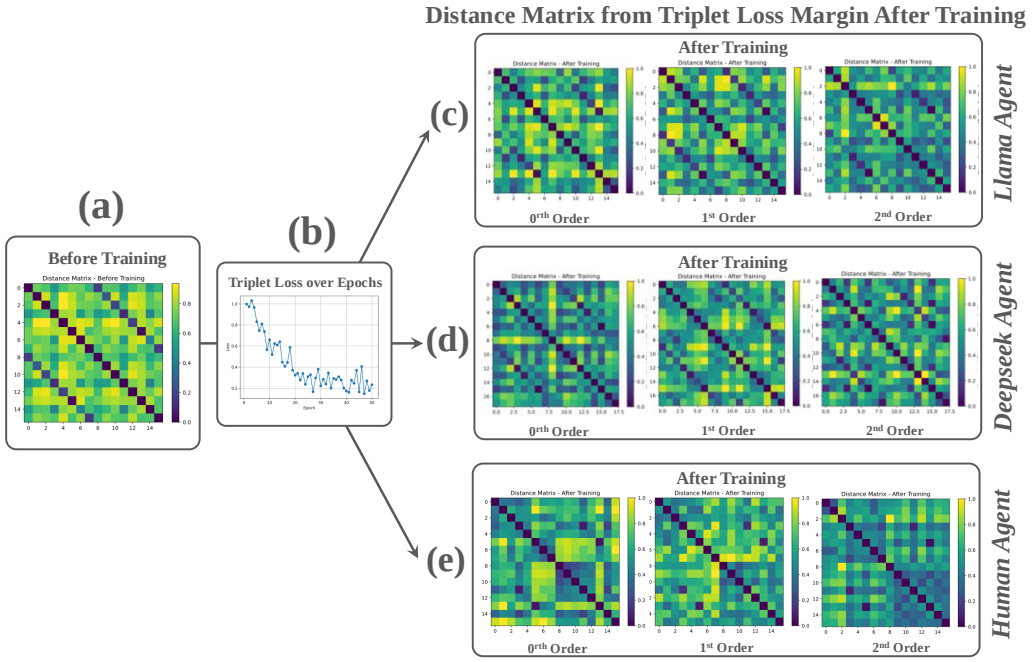


Figure 4: **Distance matrices showing representational shifts after training for different agents across ToM orders.** **(a)** Initial (pre-training) distance matrix. **(b)** Triplet Loss Over Epochs. **(c)** Llama Agent, across 0th, 1st, and 2nd-order ToM reasoning. **(d)** DeepSeek Agent across 0th, 1st, and 2nd-order ToM reasoning. **(e)** Human Agent across 0th, 1st, and 2nd-order ToM reasoning.

This is followed by a comprehensive framework for quantifying these internal representations using a triplet-based similarity approach. Each triplet, in the form of (A,B,C) where Story A is closer to Story B than to Story C, consists of an anchor (e.g., a particular story about Alice and Bob), a positive (another story deemed more similar to the anchor), and a negative (a story deemed less similar to the anchor). The framework enforces a triplet loss that encourages the learned representation of the anchor to be closer to the positive than to the negative in embedding space, thereby capturing differences in the mental states as they evolve over multiple interactions and extended time periods, enabling direct comparisons not only among AI architectures but also between AI models and human participants. This methodology makes the mental scape of a model more transparent too, illustrating how connections among beliefs, intentions, and perspectives emerge and evolve over multiple interactions and extended time periods. All in all, a view like this allows us to assess both the benefits and potential pitfalls of different architectural strategies—e.g., token-based vs. structured-reasoning models—for tasks requiring higher-order social cognition.

Our results underscore the challenges and opportunities in achieving representational alignment for Theory-of-Mind (ToM) tasks. While models such as DeepSeek and LLaMA can often answer individual ToM questions accurately, their distance matrices revealed a substantive gap between how they and humans cluster and associate False Belief stories (and other ToM tasks not presented here). Specifically, humans grouped narratives by conceptual rather than surface-level textual similarities, leading to clear block structures that reflect the underlying class distinctions (e.g., False Belief A vs. False Belief B). Current AI models, in contrast, tend to conflate stories that differ only slightly in wording but maintain distinct causal or mental-state implications. This discrepancy persisted across 0th-, 1st-, and 2nd-order ToM reasoning.

Notably, an explicit reasoning approach, as implemented in DeepSeek, did not reliably translate into improved representational alignment. Although such structured models can decompose tasks into sequential subtasks, as illustrated in the preliminary study, our triplet-based distance metrics show that they still struggle to recover the higher-level conceptual structure shared among related stories. These findings highlight that mere correctness on isolated prompts does not guarantee that a model has formed robust and coherent "mental maps" akin to those observed in human cognition.

From a practical standpoint, this work offers two key avenues for enhancing model training. First, educators or domain experts could perform test-time training by providing visual or conceptual diagrams of correct mental state mappings, thereby guiding the model toward more human-aligned representations. Second, we can train the LLMs on the embeddings of the training data rather than rote training data. Thus, we allow for the LLM to reason in an unrestricted space, where only the embedding matters, instead of a language space. This allows us another means to mold the continuous latent space representations, as seen in (Hao et al., 2024).

Thus, there remains a pressing need to systematically evaluate alignment. Quantitative measures like triplet-based distance matrices can illuminate as to how models arrive at similar or different answers, rather than focusing solely on whether or not those answers are correct. By adopting more sophisticated representational alignment metrics, researchers can pinpoint the precise ways in which a model's internal reasoning diverges from human patterns. This not only informs the design of future models aiming to handle multi-step cognitive tasks but also opens the door for specialized interventions—such as curriculum learning on targeted examples—to guide models toward conceptual structures that better match human judgments.

While our current framework focuses on static comparisons of representational similarity in Theory-of-Mind (ToM) tasks, real-world cognition—and human ToM—unfolds over time. We plan to extend our analysis into the temporal domain: how AI agents and humans update, maintain, and revise mental-state representations across sequential interactions. This includes integrating temporal smoothing or belief-tracking memory modules that model evolving cognitive maps, enabling comparisons not just at a single point but across trajectories of belief states—a closer approximation of natural social cognition. These temporal extensions also open the door to clinically meaningful applications, particularly for neurodegenerative conditions such as Alzheimer's disease. Recent fMRI-based ToM studies have identified early breakdowns in belief reasoning as potential transdiagnostic biomarkers of cognitive decline (De Lucena et al., 2020; Zegarra-Valdivia et al., 2023; Tripathi et al., 2025). These studies demonstrate that deteriorations in the ability to infer nested beliefs or social perspectives can precede more general memory-related symptoms. Our framework, which identifies coherence or fragmentation in mental-state clustering via triplet-based metrics, could offer a digital parallel to these neuroimaging-based assessments. We envision an app-based platform that monitors cognitive coherence and perspective-taking through interactive ToM narratives—non-invasive, repeatable, and sensitive to early shifts in representational structure. Embedding this into daily-use tools could help identify emerging impairments and intervene before behavioral symptoms appear. Over time, visualizing representational drift may enable clinicians to detect and track declining executive or social function, forging a new bridge between cognitive AI research and real-world neuropsychological care.

Achieving deeper alignment will also foster broader implications for AI systems that interact in social, educational, or collaborative settings. As we embed AI more deeply into real-world scenarios—ranging from tutoring platforms and social robotics to enterprise decision support—models must handle evolving, context-rich tasks involving multiple agents with interlocked beliefs. Our proposed framework not only facilitates closer scrutiny of how AI models arrive at their inferences in real time but also supplies a systematic method for aligning these inferences with human expectations. Through continued research and refinement, we envision a future in which AI systems more faithfully mirror human cognitive processes, ultimately improving both their problem-solving efficacy and their transparency in socially oriented tasks.

6 REFERENCES

Apple Machine Learning Research. (2024). *GSM-Symbolic: A Dataset and Benchmark for Symbolic Reasoning in AI.* https://machinelearning.apple.com/research/gsm-symbolic

Constantinescu, A. O., O'Reilly, J. X., & Behrens, T. E. J. (2016). Organizing conceptual knowledge in humans with a gridlike code. *Science, 352*(6292), 1464–1468. https://doi.org/10.1126/science.aaf0941

Das, R. E., & Das, R. (2024). Iterative Theory of Mind Assay of Multimodal AI Models. In *ICML 2024 Workshop on LLMs and Cognition*. https://openreview.net/forum?id=PsGVVQJZGk

De Lucena, A. T., Bhalla, R. K., Dos Santos, T. T. B. A., & Dourado, M. C. N. (2020). The relationship between theory of mind and cognition in Alzheimer's disease: A systematic review. *Journal of Clinical and Experimental Neuropsychology, 42*(3), 223–239. https://doi.org/10.1080/13803395.2019.1710112

Hao, S., Sukhbaatar, S., Su, D., Li, X., Hu, Z., Weston, J., & Tian, Y. (2024). Training Large Language Models to Reason in a Continuous Latent Space. *arXiv preprint*. https://doi.org/10.48550/arXiv.2412.06769

Jamieson, K. G., & Nowak, R. D. (2011). Low-dimensional embedding using adaptively selected ordinal data. In *Allerton Conference on Communication, Control, and Computing*.

Messeri, L., & Crockett, M. J. (2024). Artificial intelligence and illusions of understanding in scientific research. *Nature, 627*, 2024. https://doi.org/10.1038/s41586-024-07146-0

Mitchell, M. (2024). Debates on the nature of artificial general intelligence. *Science, 383*(6689). https://doi.org/10.1126/science.ado7069

Mitchell, M., & Krakauer, D. C. (2023). The debate over understanding in AI's large language models. *Proceedings of the National Academy of Sciences, 120*(13). https://doi.org/10.1073/pnas.2215907120

Mitchell, M., Palmarini, A. B., & Moskvichev, A. (2023). Comparing humans, GPT-4, and GPT-4V on abstraction and reasoning tasks. *arXiv preprint*. https://doi.org/10.48550/arXiv.2311.09247

Peterson, J. C., Abbott, J. T., & Griffiths, T. L. (2018). Evaluating (and improving) the correspondence between deep neural networks and human representations. *Cognitive Science, 42*(8), 2648–2669. https://doi.org/10.1111/cogs.12655

Sclar, M., Yu, J., Fazel-Zarandi, M., Tsvetkov, Y., Bisk, Y., Choi, Y., & Celikyilmaz, A. (2024). Explore Theory-of-Mind: Program-Guided Adversarial Data Generation for Theory of Mind Reasoning. *arXiv preprint*. https://doi.org/10.48550/arXiv.2412.12175

Shepard, R. N. (1980). Multidimensional scaling, tree-fitting, and clustering. *Science, 210*(4468), 390–398. https://doi.org/10.1126/science.210.4468.390

Strachan, J. W. A., Albergo, D., Borghini, G., Pansardi, O., Scaliti, E., Gupta, S., Saxon, K., Rufo, A., Panzeri, S., Manzi, G., Graziani, M., & Becchio, C. (2024). Testing theory of mind in large language models and humans. *Nature Human Behaviour, 8*(7), 1285–1295. https://doi.org/10.1038/s41562-024-01882-z

Sucholutsky, I., & Griffiths, T. L. (2023). Alignment with human representations supports robust few-shot learning. *arXiv preprint, arXiv:2301.11990*. https://doi.org/10.48550/arXiv.2301.11990

Tripathi, V., Batta, I., Zamani, A., Atad, D. A., Sheth, S. K. S., Zhang, J., Wager, T. D., Whitfield-Gabrieli, S., Uddin, L. Q., Prakash, R. S., & Bauer, C. C. C. (2025). Default Mode Network Functional Connectivity As a Transdiagnostic Biomarker of Cognitive Function. *Biological Psychiatry: Cognitive Neuroscience and Neuroimaging, 10*(4), 359–368. https://doi.org/10.1016/j.bpsc.2024.12.016

Wang, Q., Walsh, S., Si, M., Kephart, J., Weisz, J., & Noel, A. (2024). Theory of mind in human–AI interaction. In *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*. https://doi.org/10.1145/3613905.3636308

Zegarra-Valdivia, J. A., Rijpma, M. G., Shany-Ur, T., Kramer, J. H., Miller, B. L., & Rankin, K. P. (2023). Cognitive and emotional theory of mind in dementia: Impact on real life behaviors. *Alzheimer's & Dementia, 19*(S4), e067855. https://doi.org/10.1002/alz.067855

Zhang, W., Qian, Y., Cao, X., & Shi, S. (2024). Evaluating Theory-of-Mind in Large Language Models: A Case Study on Second-Order Belief Reasoning. *arXiv preprint*. https://doi.org/10.48550/arXiv.2412.06769

OpenAI ChatGPT_o1. LLM Conversation Snippets. https://chatgpt.com/share/67a3e5bc-1444-800b-ba8f-15e2bd1e22c7

OpenAI ChatGPT_4o. LLM Conversation Snippets. https://chatgpt.com/share/67a3e966-66e0-800b-9ed7-c354d1fd5315

OpenAI ChatGPT_4. LLM Conversation Snippets. https://chatgpt.com/share/67a3e984-e10c-800b-a492-0ded61176d63