

ToMCAT: Benchmark for Socially Assistive Robots with Theory of Mind of Children Assembling Tangram Puzzles

Jason R. Wilson¹, Irina Rabkina², Mark Roberts³ and Laura M. Hiatt³

Abstract—Assistive robots will be more effective if they can accurately reason about the intentions and beliefs of the user (i.e., have Theory of Mind (ToM)). ToM benchmarks allow us to examine how well an artificial agent (e.g., robot) is able to do ToM reasoning in a given scenario. However, there is a need for ToM benchmarks that are more representative of the challenges faced in assistive robotics. Existing benchmarks from AI and HRI make simplifying assumptions, such as simply defined goals, plans that are indicative of goals, and no user errors. To address the challenges from relaxing these assumptions, we propose the Theory of Mind of Children Assembling Tangrams (ToMCAT) dataset. The data is derived from videos of children building tangram puzzles while being assisted by a social robot. As a baseline benchmark, we evaluated two approaches for how well they can recognize which puzzle this child is building based on a single observation. Analogical reasoning correctly recognized the puzzle more than 75% of the time and had perfect accuracy for puzzle states that were close to complete. However, an out-of-the-box commercial LLM correctly recognized the puzzle only 60% of the time and was accurate on less than 80% of the completed puzzles. Our results suggest that the ToMCAT dataset offers challenges for recognizing the intended puzzle of a child. Furthermore, the dataset provides opportunities to examine additional ToM reasoning capabilities. Overall, the ToMCAT dataset provides a useful benchmark to facilitate the advancement of ToM reasoning for assistive robotics.

I. INTRODUCTION

In situations where assistive robots are tasked with helping a human complete a task, it is beneficial and often necessary to have the robot understand what the human is trying to accomplish (i.e., understand the human’s goal), predict what they will do next, and infer what assistance would help them progress in the task. Each of these activities is central to classic Theory of Mind (ToM) challenges. ToM is the capability to understand that others can hold differing beliefs, knowledge, desires, or intentions [1]. ToM reasoning, in turn, is the inference of any one (or more) of these aspects of cognition in another agent.

To evaluate how well an artificial agent can use ToM reasoning, there is a growing interest in community datasets that can provide benchmarks of ToM performance (e.g., [2], [3], [4], [5], [6]). These benchmarks help to standardize and

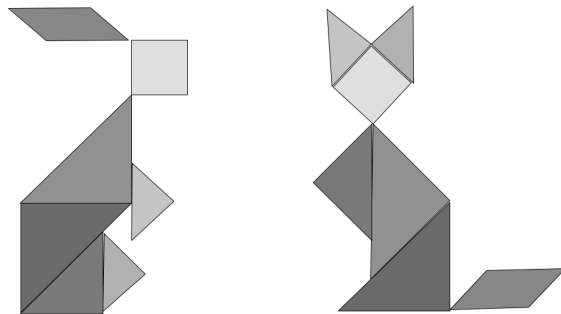


Fig. 1. Illustrations of two Tangram puzzles using the same individual pieces, but forming different animals: a rabbit (left) and cat (right).

compare ToM approaches. However, the majority of datasets underlying these benchmarks make one (or more) of several simplifying assumptions. First, many assume that the goal of the agent can be represented as a single expression (e.g., *inventory_contains(bread)* for a Minecraft-like domain, or *at(truck1)* for a domain involving predicting what food truck an agent is going to). In addition, these domains are often restricted to include predictable and/or legible plans. A predictable plan reduces ambiguity by having few possible plans for a goal, and a legible plan reduces ambiguity by having few goals that correspond to a given plan [7]. For example, walking the only shortest path to a food truck is a predictable plan, while a robot reaching in a direction and has only one reachable object in the direction is a legible plan. Conversely, many real-world tasks that an assistive robot would be expected to reason about are neither predictable nor legible.

We make two main contributions. First, we present a new ToM dataset in the domain of Tangram puzzles, Theory of Mind of Children Assembling Tangrams (ToMCAT), that pushes the boundary of these prior limitations in several ways. Tangram puzzles are puzzles that are solved by arranging different baseline polygons to form a larger, target shape; two such puzzles are shown in Fig. 1. While assembling such puzzles, there is a very high number of possible paths to completion: the user can place any one of 7 pieces in any orientation, and in any relation to pieces that are already on the board. Goals (i.e., target shapes) must also be described using multiple expressions, denoting the spatial relationships of adjacent pieces. Furthermore, users (in this dataset, children) often make mistakes, making it even harder to infer which target shape they are assembling before the puzzle is complete. All of this means that ToMCAT offers a ToM task that is more similar to the kinds of tasks assistive

JRW was supported by the National Science Foundation under Grant No. 2338148; MR and LMH thank NRL for supporting this research.

¹Franklin and Marshall College, Lancaster, PA 17604, USA
jrw@fandm.edu

²Barnstorm Research Corp., Malden, MA 02148, USA
irina.rabkina@barnstormresearch.com

³Navy Center for Applied Research in Artificial Intelligence,
US Naval Research Laboratory, Washington, DC 20375,
USA
mark.c.roberts20.civ@us.navy.mil,
laura.m.hiatt.civ@us.navy.mil

robots are likely to perform in the real world.

Second, we evaluate two techniques—LLM reasoning and analogical ToM—on the task of inferring which Tangram puzzle a child is assembling (cf. Fig. I). Recent advances in LLM technology suggest that LLMs can perform spatial reasoning [8] and work well for various ToM tasks (e.g., [2], [9]). Similarly, analogical reasoning has been shown to effectively reason about both spatial relationships [10] and ToM [11]. Our results indicate that recognizing the child’s goal is challenging for an LLM, but relatively straightforward for the analogical approach. The evaluation demonstrates that the ToMCAT dataset offers a non-trivial ToM task, as well as provides initial baselines for future comparison of ToM applied to ToMCAT.

II. BACKGROUND ON TOM BENCHMARKS

While ToM benchmarks have standardized the evaluation of ToM approaches, many benchmarks make several simplifying assumptions that limit how well they evaluate the ToM reasoning abilities of assistive robots.

A. Goal Complexity

Often, ToM benchmarks are framed as goal recognition problems, with the reasoner’s main inference being the goal state of the observed agent. Unlike other aspects of ToM (i.e., desires, beliefs, etc.), goals are observable, and thus can be extracted from the dataset as ground truth. In fact, even when the stated task is to infer a different aspect of ToM, evaluation is typically based on the goal state (e.g., the food truck domain [12], where stopping at the location of a given food truck serves as a proxy for the user’s preference).

Furthermore, goal representation is sometimes simplistic. For example, in stag-hunt [13], agent goals, such as the identity of the hunting target or the intent to collaborate, are represented as a single expression or grid coordinate pair. Similarly, the goal in the food truck domain is often represented as location coordinates [12]. While goal states simplify the inference task, they may not require reasoning about the entire scene or, crucially, the internal state(s) of the observed agent(s).

B. Plan Predictability and Legibility

Many ToM benchmarks also make simplifying assumptions about how goals are achieved. For example, the food truck domain [12] assumes that an agent: has few possible paths to its goal from its current location, always behaves correctly, and its destination maps directly to their goal. Many ToM domains make similar assumptions, allowing straightforward inference of agent intentions solely from movement in a grid world (e.g., [13], [14], [15], [16]; see [17] for a detailed list).

Near-perfect predictability of goals from observed movements significantly simplifies the problem of ToM reasoning, and is not reflective of the kinds of problems assistive robots are likely to face in the real world. In fact, full ToM reasoning requires a deeper understanding of the internal states motivating an agent’s actions, rather than focusing on observable motion trajectories.

C. Noisy Data and User Errors

In addition to being predictive, observations are usually assumed to be reliable. The observer is unobstructed in their observations of the other agent (e.g., moving a book to another room [2] or reaching for an object [18]). Furthermore, it is generally assumed that the agent’s actions are always purposeful and correct, as there is no reason to believe that it has imperfect information or does not know what it is doing.

There are several exceptions to this. For example, in one study, a social robot is tasked with recognizing the reason a human teammate took an incorrect turn and offering assistance [19]. Similarly, false belief benchmarks (e.g., [3], [20], [21]) assume that the observed agent is missing important information and might therefore make mistakes. However, these benchmarks are typically represented as stories, and do not directly translate to ToM reasoning on the fly.

The Minecraft dataset presented in [22] and [6] explicitly introduces noise into observations of an agent performing a crafting task in the game. There are inherently multiple ways to accomplish each possible task (e.g., ingredients can be collected in different orders) and the agent may skip items because it is already in their inventory. Performing ToM in this dataset therefore requires reasoning about both the agent’s movements and knowledge about the task. Yet, the dataset is primarily a goal recognition task, with goals represented via a single expression (e.g., *in_inventory(bread)*), thus still making simplifying assumptions.

In contrast, a Tangram puzzle possesses complexity along each of the three axes listed above. A puzzle’s goal, given to human solvers as the visual depiction in Fig. I, are computationally represented as a (sometimes large) set of states that specify the relationship of all of the individual target pieces to one another. They also have many possible paths to the successful completion of the goal, since pieces can be placed in any order. And, finally, people typically make many mistakes when performing these puzzles; this is especially true for children, as we will see below.

III. TANGRAM PUZZLES: THE PROBLEM DOMAIN

Tangram puzzles are solved by arranging different baseline polygon-shaped pieces to form a larger, target shape (Fig. I). There are 7 pieces to each puzzle: 2 large triangles, 2 small triangles, and 1 each of a medium triangle, square, and parallelogram. Tangram puzzles may be employed to teach geometry to young children [23]. Since spatial reasoning skills have been shown to support the development of early mathematics skills [24], spatial assembly tasks, like tangram puzzles, are used to assess early geometric and spatial reasoning in young children [25].

Despite traditionally being a child’s puzzle, intricacies of these puzzles contribute to some challenges for effective ToM reasoning. Although tangram puzzles are inherently visual, it is very challenging to infer a puzzle solver’s goal until the puzzle is nearly complete. This is partially true because of the redundancy of shape placement and orientation that occurs in many tangram solutions. More so, it is challenging because

of the adjustments puzzle solvers employ while assembly the puzzles, both to fix errors previously made, as well as to fine-tune the alignment of shapes relative to one another. When the adjustments are incorrect, understanding the child’s intent is especially challenging because the purpose of their actions becomes ambiguous.

Tangram puzzles can be challenging Theory of Mind problems for an observer of the child assembling the puzzle pieces. For an observer who is assisting the child, the observer would at least need to know what puzzle the child is intending to build. The observer may use this knowledge to determine how to best assist the child, which may require understanding what part of the puzzle the child is working on. When a child makes a mistake and misplaces a piece, if the observer were to infer why the child believed that the piece positioning was correct, the assistance could help the child correct their understanding instead of simply telling them what needs to be fixed.

IV. TANGRAM BENCHMARK


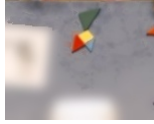
We propose a new benchmark measuring a robot’s ability to reason about a user’s task. The benchmark uses a new Theory of Mind of Children Assembling Tangrams (ToMCAT) dataset¹, which consists of observations of a child assembling tangram puzzles with the assistance of a socially assistive robot. The child is given one of two illustrations (see Fig. 1) of a completed puzzle and is instructed to build the puzzle shown in the illustration.

The dataset represents a situation in which the child and robot are using knowledge of the task but the knowledge is not necessarily shared. The robot knows how the tangram pieces fit together to create each of the completed puzzles. Since the robot also knows the size of each piece, it can reason about the interchangeability of pieces of the same size. While the child has an illustration of the puzzle, that illustration is in black and white, which often results in insufficient knowledge about each piece and how the pieces need to relate to each other to build the puzzle. For example, the red medium triangle often causes confusion. Many children try to use the triangle where a larger or smaller triangle should be used. Intermediate states of the puzzle as it is assembled, interchangeable pieces, and naturally occurring errors are features that other tangram-related datasets (e.g., KiloGram [26]) do not have.

In the current work, we use this dataset to benchmark how well AI reasoners can recognize which puzzle the child is assembling. Recognition of the puzzle would then enable the robot to make further inferences requiring ToM, such as inferring which part of the puzzle the child is working on, what the child may do next, and any misconceptions the child has that interfere with their ability to complete the task. Although the dataset we provide supports benchmarking these other reasoning steps, we do not currently assess how AI reasoners perform on these subsequent processes.

¹<https://github.com/FandM-CARES/ToMCAT>

TABLE I
SAMPLES FROM TOMCAT DATASET

ID	Time	Shape	Screenshot	Relations	+	-
s02	3:55	rabbit		(vc sva pvc1) (vc svd gva) (teq (EdgeFn ovb ovc) (EdgeFn gvb gvc))	3	0
s36	6:21	cat		(teq (EdgeFn sva svb) (EdgeFn bvb bvc)) (tovi (EdgeFn svb svc) (EdgeFn rva rvb) svc rva) (to (EdgeFn bva bvb) (EdgeFn rva rvb) bvb rvb) (vc svd gva)	2	2

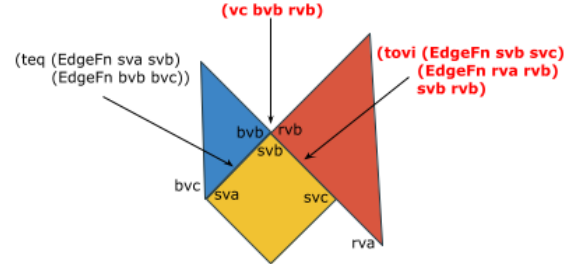


Fig. 2. An illustration of correct and incorrect relationships between pieces. The intended puzzle is a cat, and a medium triangle is used in place of a small one. This results in 1 correct relationship between the blue triangle and the square. The other two relationships are incorrect.

A. Data Collection

The dataset is based on videos of child-robot interactions collected as part of a previous study [27]. Each participant ($N = 34$), aged 5 to 8 years old, assembled either a rabbit or cat puzzle with the help of a Misty robot. The remotely controlled robot provided social assistance to the child by providing feedback and encouragement [28]. The videos are from a camera mounted on the wall above the robot, providing a top-down view of the puzzle that is not disturbed by the movement of the robot. For each interaction, we reviewed the videos to identify when the state of the puzzle has changed and the child is not still moving pieces of the puzzle. Each new state of the puzzle was then recorded as a new observation (see Table I for samples from the dataset). The observed puzzle state was recorded in two forms: (1) A single frame of the video was transformed to account for projection and cropped to isolate the portion of the image with the puzzle. (2) A symbolic representation of the relationships between the puzzle pieces (cf. next section for details on the representation). The symbolic representation was manually encoded, and to ensure accuracy, each video was reviewed by at least 2 annotators.

We also annotated each observation with two measures of puzzle correctness: (1) The number of correct puzzle relationships in the observation (2) The number of incorrect puzzle relationships. Fig. 2 shows an example from the dataset that has 1 correct relationship and 2 incorrect ones. In Table I, the + and - columns represent the correct and incorrect relationships counts, respectively.

B. Puzzle Representation

To represent the structure of the puzzle, we developed a novel representation designed to model spatial adjacency relationships between 2D, non-overlapping, convex polygons. The representation is an extension of the Region Connection Calculus and expands upon the types of relationships defined in the Externally Connected (EC) relation [29], [30]. Polygons are represented as a series of vertices, and each edge is a pair of vertices. Our representation uses these edges and vertices to describe all 8 possible relations between adjacent edges and vertices (see Table II). Since the set of representations are mutually exclusive and collectively exhaustive, we can safely assume that any undefined relation implies that the polygons are not adjacent.

The picture of the target puzzle can be succinctly represented as a set of relationships between adjacent pieces. This condensed representation abstracts the problem away from the pixels available in each frame of a video and disentangles the robot’s reasoning about the puzzle from the complexities of the robot’s vision system.

C. Task and Dataset Analysis

The tangram task provides a small (7-piece), well-defined task, but it also involves a relatively high degree of complexity. Both the cat and rabbit have 8 relationships between pieces in a correctly completed puzzle. Considering the interchangeability of some of the pieces, there are 4 correct arrangements of the pieces for each target puzzle, resulting in a total of 8 possible goal solutions.

Since the pieces can be assembled in any order, there are 7! optimal plans to place the pieces to complete the task. The complexity of the problem is also related to the large number of ways in which pieces may be related to each other. Since pieces can be related through each of their sides and vertices using the 8 predicates defined in Table II, there are over 900 possible relationships between all of the puzzle pieces.

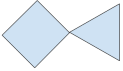
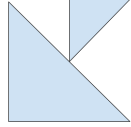

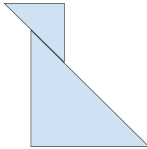

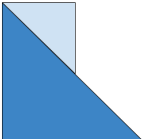
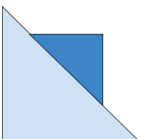
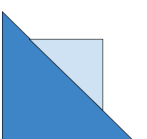
Some of the complexities of the task are apparent in our dataset. The dataset consists of 34 children assembling either a cat or rabbit puzzle (15 cat, 19 rabbit). There are 436 observations in total (avg 12.8 obs/child). The quickest puzzle assembly resulted in 6 observed state changes, and the slowest puzzle assembly had 42 observations.

The children took an average of nearly 13 steps to assemble 7 pieces; many children made suboptimal choices. Furthermore, children tended to improve the puzzle more the closer it got to completion. To start the puzzle, only 47% (16/34) of the children correctly paired the first two pieces. They continued to have difficulty in the early stages and made improvements to the puzzle only 49% of the time while the puzzle had 2 or fewer correct relations. With 3 to 5 correct relations in the puzzle, children improved the puzzle 58% of the time, and with 6 or more correct relations, they improved the puzzle 70% of the time.

V. EVALUATION

Our evaluation measures how well two approaches can determine which puzzle the child is building. We compared

TABLE II
POLYGON ADJACENCY RELATIONSHIPS. DARKER POLYGONS ARE THE POINT OF REFERENCE AND ARE DEFINED FIRST IN THE RELATION.

Name	Representation	Example Image
Vertex Connection	$(vc\ v_s\ v_t)$	
Vertex Edge Connection	$(vec\ v_s\ e_t)$	
Tangential Equal	$(teq\ e_s\ e_t)$	
Tangential Overlapping	$(to\ e_s\ e_l\ v_s\ v_l)$	
Tangential One Vertex Included	$(tovi\ e_s\ e_l\ v_s\ v_l)$	
Tangential One Vertex Included Inverse	$(tovii\ e_l\ e_s\ v_l\ v_s)$	
Tangential Edge Included	$(tei\ e_s\ e_l)$	
Tangential Edge Included Inverse	$(teii\ e_l\ e_s)$	

analogy approach as well as a commercial LLM performs on the same task. The performance of each approach is presented as a baseline for future work to build upon.

A. ANALOGY FOR THEORY OF MIND

Analogical reasoning is a cognitive process by which previous knowledge is compared to the current problem in order to apply prior decision making, inferences, conclusions, etc. to the current situation. In addition to being a key aspect of human reasoning, analogy has a wide range of applications in robotics and computational systems. It has been used to generalize human input in human-robot-interaction scenarios,

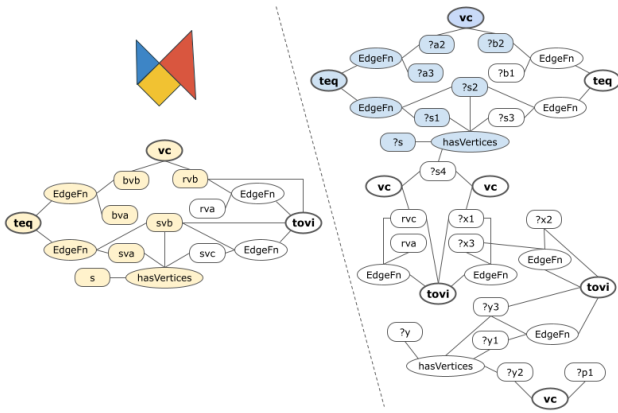


Fig. 3. Diagrams of mappings between structures. The left is an illustration of 3 pieces and the corresponding representation displayed as a graph. The right is a representation of the cat solution. The shaded elements represent the elements that are mapped to each other. Some elements are not shown since the full representation is too large to show here.

[31], [32], amplifying costly human demonstrations; it has also been successfully applied to geometric reasoning [10], goal reasoning [6], and theory of mind [11]. Given these prior successes, especially on theory of mind tasks, we investigate here how well analogy infers the child’s target puzzle for the ToMCAT dataset.

In order to determine what prior knowledge applies to the current situation, analogical reasoning first needs to align any relevant prior scenarios with the current one. The primary step in this alignment is finding elements (e.g., expressions, relations, attributes, or patterns thereof) that are similar in the two scenarios and creating a mapping between them. The more mappings that there are between scenarios, the higher the scenarios’ similarity [33]. In Fig 3, the 11 shaded elements are mapped because they share similar structures. This alignment process serves two purposes: (1) it allows reasoning about scenarios that don’t have the exact same representation as the current scenario; and (2) it allows the reasoning process to identify the most similar scenario(s) to the current one. Any underspecified elements in the current scenario, such as what action to take next, can then be inferred from the corresponding element in the prior scenario. Here, we used the Structure Mapping Engine (SME)’s implementation of this process. [34].

In the tangram task, the current scenario (i.e., the current state of the puzzle assembly) was represented as expressions specifying the attributes of each piece (e.g., color, shape, relative size), as well as expressions specifying the relationships of the shapes’ positions, as discussed in Section IV-B. The puzzle’s final shape (cat and rabbit) was similarly represented, but did not specify colors for the redundant triangles in order to encompass all possible solutions.

When using analogical reasoning to infer the child’s target puzzle, each target goal (i.e., cat, rabbit) was compared to the current scenario. The goal that had greater analogical similarity with the current scenario was selected as the inferred puzzle solution; intuitively, this was the goal with the most

interconnected pieces in common with the current scenario. Each scenario was tested independently; prior scenarios were not taken into account when making each inference.

B. LLM Comparison

Since the use of LLMs has rapidly become a common approach for implementing advanced reasoning capabilities in social robots, we compared the performance of analogy to a commercial LLM. Our evaluation used the API interface to Claude Sonnet 3.5, with no fine-tuning. For a straightforward comparison with our analogy-based approach, we used the same logical representation of each puzzle and observation state. We chose not to convert these representations to natural language because the logical representation was more concise and did not risk having any ambiguities inherent to using natural language.

The initial prompt to the LLM consisted of three parts:

- A description of the problem: “We are trying to infer the puzzle that a person is building based on how the pieces are connected to each other. In order to communicate these connections, we are using first-order logic to represent the relationship between tangram puzzle pieces.”
- A natural language description of each of the predicates used in the logical description.
- A description of the LLM’s task: “Your task is to interpret whether the participant is making a cat or rabbit puzzle based on the given series of connection. Only answer Cat or Rabbit, EVEN IF YOU CAN’T INTERPRET THE REPRESENTATION.”

For each observation state, we include the prompt “Which puzzle is being built if the current arrangement of pieces is the following:”, followed by the the same logical description of the observation state that is given to analogy.

Using a temperature setting of 0.8, we found some variability in the responses from the LLM. As a result, we made 3 independent requests to the LLM and used a best of 3 voting scheme to determine the model’s final inference. As with analogy, each observation state is tested independently.

C. Results

The overall accuracy of analogy was 75.2%, and the accuracy of the LLM was 60.1%.

To further investigate how well each algorithm performs, we analyze their accuracy relative to the number of correct and incorrect expressions in the puzzle. A completed puzzle has 8 correct expressions and no incorrect expressions. An expressions is considered correct if it can be found in a solution to the puzzle. For interchangeable pieces (e.g., the small and large triangles), we consider the expression to be correct if it involves a piece of the correct size, regardless of the color. For example, for the cat, the blue and pink triangles that form the ears can be on either side of the yellow square.

For every observed state, we counted the number of correct and incorrect expressions currently found in the puzzle. Then for each combination of correct and incorrect counts, we calculated the accuracy of each algorithm. Fig. 4

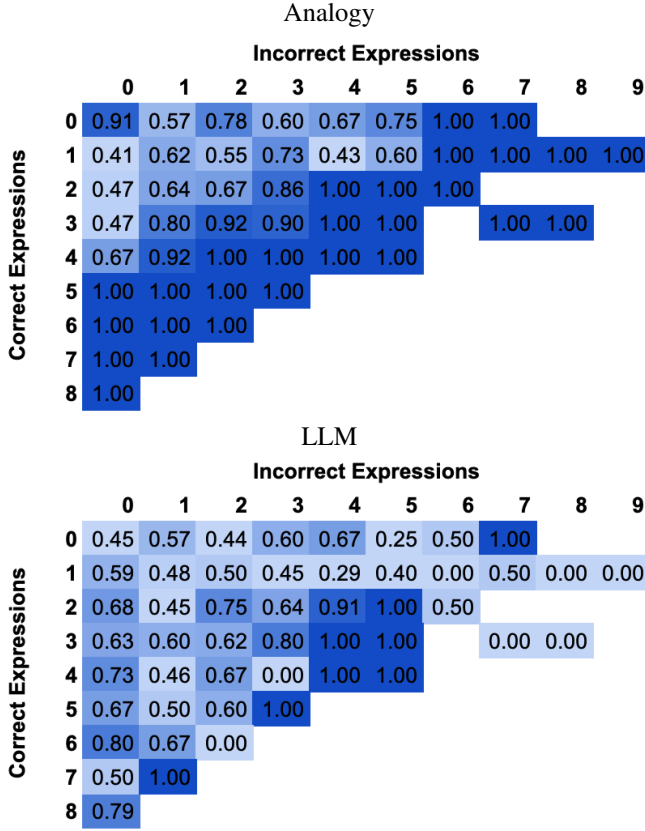


Fig. 4. Heatmaps showing the accuracy of the algorithm for each combination of correct and incorrect relations, where darker colors indicate more correctness. Positions with no color and value indicate that the dataset did not have an example of that combination of correct/incorrect relations.

shows heatmaps based on these accuracies. Each combination containing 5 or more incorrect expressions happened 4 or fewer times across the dataset, so we draw limited generalization from this small sample; they are displayed only for completeness. In contrast, almost every combination with 1 or fewer incorrect relations happened 6 to 29 times.

VI. DISCUSSION

Robot assistance often requires the robot to reason about the task being performed by the user. The robot may use knowledge it has about the task to infer what the user is attempting to do, what they may do next, and how to help them complete the task. We consider scenarios where the child is a user, and the robot is expected to have greater knowledge of the task than the user. Using Theory of Mind (ToM), the robot should be able to distinguish between its knowledge of the task and the child’s knowledge and use any differences to help the child in completing the task. The first step in this process is recognizing what the user is doing. The paper provides two contributions: a new dataset with two baseline performance evaluations, which assess how well a robot using these approaches would recognize TomCAT.

A. Tangram Dataset

The ToMCAT dataset serves as a new ToM benchmark for the HRI research community. This dataset captures 436 observations of 34 children being assisted by a socially assistive robot while building a tangram puzzle. The dataset addresses each of the three limitations from Sec. II.

1) *Goal Complexity*: If we view the puzzle solution as the “goal”, then a goal in this task is a correct arrangement of the pieces. Instead of a single expression defining each goal, both the cat and rabbit have 8 relationships between pieces in a correctly completed puzzle. Additionally, each goal can have 4 possible solutions, resulting in 8 arrangements of pieces as possible goals. The complexity of each goal in sharp contrast to much of the AI research on ToM that uses single-expression, destination-oriented goals (e.g., *at(truck1)*). Complex goals are likely with assistive robotics, where we see goals like correctly assembling furniture [35], sorting medications [36], or completing an exercise routine [37].

2) *Plan Predictability and Legibility*: The tangram task is small but solving the puzzles can be complex. There are 7! optimal plans for solving the puzzle, and an infinite number of suboptimal plans. For the first step in the plan, there are over 900 possibilities as a result of the many orientations and positions a piece can have relative to another piece. Given the many possibilities, many of the children made several mistakes in the earlier stages of the task. Since some children were able to solve the puzzle optimally, it suggests that the complexity of the problem can be managed through application of spatial and geometric reasoning to create heuristics and constraints to guide the search through this large problem space.

3) *User Errors*: The children in the ToMCAT dataset often made mistakes, which presents new challenges in determining what the child’s intentions were. When a piece is misplaced, it may not be clear that the child made an error. Instead, the child may simply be making a different puzzle. For example, in the sixth observed state for one participant, the child connected the two large triangles in a way that it would correctly form the rabbit’s body. This could lead the reasoner to believe that the child is building a rabbit. However, the child was actually trying to build the cat, and the connection between the triangles was made in error. In almost every ToM dataset and evaluation we have found, it is assumed that the observing agent knows what they are doing, is making rational choices, and is capable of acting as intended. However, in assistive robotics, one or more of these may not be true, and the robot needs its ToM reasoning to be robust to the ambiguities created by user errors.

B. Evaluation of Analogy and LLM on TomCAT benchmarks

To provide a baseline benchmark, we evaluated how well analogical reasoning and one commercial LLM are able to recognize which puzzle is being assembled by the child based on a single observed state. As expected, there is considerable room for improvement for both approaches. A clear opportunity for improvement would be using a *sequence* of observations. This could be particularly effective when

the child made a mistake. For example, one child’s puzzle had 6 correction relationships and no incorrect relationships between the puzzle pieces, and then they started making mistakes. Soon the puzzle had 3 correct and 1 incorrect relationships, which caused the LLM to incorrectly recognize the puzzle. After some more tinkering, the puzzle had 3 correct and 3 incorrect relationships, which caused both models to incorrectly recognize the puzzle. Future work should consider how to integrate these previous observations into the reasoning process.

Even when there are no mistakes in the puzzle, there is space for considerable improvements in how an LLM reasons about this problem. A completed puzzle with 8 correct relations and 0 incorrect relations was accurately recognized only 79% of the time. While this points to the LLM not performing the task well, we note that this is off-the-shelf performance. It is likely that fine-tuning and/or advanced prompting techniques would improve the LLM’s understanding of the task. However, the goal of the present work was to set baseline benchmarks, rather than identify the best algorithms for the task. We leave the latter as a challenge to the community.

In general, we would expect a reasoner to more accurately reason about the puzzle as it gets closer to completion. Children in our dataset were nearly perfect in positioning the final pieces of the puzzle once there were 6 correct expressions. They are able to take correct actions at this point because they can see how the arrangement of their puzzle pieces resemble the picture of the puzzle. As a result, we expect an AI reasoner to always recognize a completed problem, and even puzzles that are nearly complete should also be recognized with very high accuracy. We see this behavior in the analogical inferences. When there are 5 or more correct expressions, analogy recognizes the puzzle with perfect accuracy, regardless of how many incorrect relations are present. Generally, we see the accuracy of analogy consistently improve once there are at least 2 correct expressions. Once there are 2 correct expressions, there are larger structures that may be mapped, leading to greater similarity.

C. Beyond Puzzle Recognition

The TomCAT dataset provides clear opportunities to investigate ToM reasoning beyond puzzle recognition. A common mistake made by the children helps illustrate this. Multiple children used the red medium triangle in place of either a smaller or bigger triangle. In particular, some children when building the cat puzzle would use the medium triangle as one of its ears instead of small triangles for each of its ears (see Fig. 2). This common confusion is likely caused by a few factors, including the puzzle illustration being black and white, making it unclear where the red triangle should go, and one of the small triangles was initially hidden but the cat clearly needed two triangles for its ears. When a child uses a medium triangle in place of a small one, there is an opportunity for the reasoner to compare which piece is used to make the cat’s second ear (using its own knowledge) to

which piece the child has put in that position (which is an indicator of the child’s beliefs). Along with knowing that a medium triangle is incompatible with a small triangle, the ToM reasoner would be able to infer that the child has a false belief regarding how the medium triangle can be used. Furthermore, puzzle recognition could be further facilitated by using counterfactual reasoning in this case to replace the medium triangle with one of the small triangles. If the counterfactual case is a much better match to the puzzle solution, then the change made for the counterfactual could be the suggested change that the robot makes to the child. Since the dataset includes many errors and most datasets include little to no errors made by the observed agent, our dataset is particularly well-suited for developing ToM reasoning to recognize and recover from observed errors.

VII. LIMITATIONS

One limitation of the TomCAT dataset at present is that this paper evaluates two puzzles—rabbit or cat—resulting in a binary classification task. In fact, one might argue that it is primarily a visual recognition problem. However, pilot experiments using Claude Sonnet 3.7 on video frames from which puzzle states were extracted for the dataset yield chance performance. There are two possible reasons for this: (1) image quality is not sufficient for VLLMs to recognize the puzzle or (2) the puzzle requires reasoning beyond pure visual recognition. We believe it is a combination of the two, but the exact balance will need to be empirically determined.

Each observation in TomCAT is based on manually extracting puzzle relationships from the video. The data is a clean and accurate representation of the actual puzzle arrangements. Also, the logical representation is well suited for analogical reasoning, which accounts for some of the higher performance of that approach. However, we believe that this representation can be computationally derived from the video. The processing of the videos is not expected to be perfect, resulting in noisy data. Given that our results here show that analogy has a high level of accuracy when there are incorrect relations (which can be viewed as similar to having incorrect data) and prior research showing analogy being robust to unreliable data [6], it is likely that analogy will continue to accurately recognize the puzzle when operating on data that is not manually constructed.

VIII. FUTURE WORK

The evaluations we presented here are intended to serve as a baseline, and there are many opportunities to improve upon these algorithms. One opportunity is for the algorithms to consider previous observations instead of reasoning about each observation independently. From a sequence of observations, which is more commonly used in ToM scenarios, the reasoner may be able to identify patterns that would make the intentions of the child more clear. Sequences of observations would also facilitate predicting and evaluating the next change that the child makes to the puzzle.

Future developments of this dataset will support further ToM benchmarking. We are continuing to review the dataset

to label states for false beliefs the child has (e.g., the medium triangle can be used in place of a small one). The videos also capture the robot’s interventions, thus providing an opportunity to link everything from recognizing the puzzle, to evaluating the child’s action, to assessing any false beliefs, to finally deciding what the robot will do to intervene.

IX. CONCLUSION

We present the ToMCAT benchmark to evaluate Theory of Mind capabilities when reasoning about children’s goals in tangram puzzles. The benchmark introduces key challenges as a result of incorporating greater goal complexity, plans with reduced predictability and legibility, and frequent user errors. We provide baseline measurements, with analogy performing perfectly on puzzles at least 50% complete.

ACKNOWLEDGMENT

We greatly appreciate Zach Locher for his contributions in building the dataset.

REFERENCES

- [1] A. M. Leslie, “Pretense and representation: The origins of” theory of mind,” *Psychological review*, vol. 94, no. 4, p. 412, 1987.
- [2] H. Shi, S. Ye, X. Fang, C. Jin, L. Isik, Y.-L. Kuo, and T. Shu, “Mumatom: Multi-modal multi-agent theory of mind,” in *Proceedings of the 39th Annual AAAI Conference on Artificial Intelligence*, 2025. [Online]. Available: <https://arxiv.org/abs/2408.12574>
- [3] H. Kim, M. Sclar, X. Zhou, R. L. Bras, G. Kim, Y. Choi, and M. Sap, “Fantom: A benchmark for stress-testing machine theory of mind in interactions,” *arXiv preprint arXiv:2310.15421*, 2023.
- [4] H. Xu, R. Zhao, L. Zhu, J. Du, and Y. He, “Opentom: A comprehensive benchmark for evaluating theory-of-mind reasoning capabilities of large language models,” *arXiv preprint arXiv:2402.06044*, 2024.
- [5] Y. He, Y. Wu, Y. Jia, R. Mihalcea, Y. Chen, and N. Deng, “Hi-tom: A benchmark for evaluating higher-order theory of mind reasoning in large language models,” *arXiv preprint arXiv:2310.16755*, 2023.
- [6] I. Rabkina, P. Kantharaju, J. R. Wilson, M. Roberts, and L. M. Hiatt, “Evaluation of goal recognition systems on unreliable data and uninspectable agents,” *Frontiers in Artificial Intelligence*, vol. 4, p. 734521, 2022.
- [7] T. Chakraborti, A. Kulkarni, S. Sreedharan, D. E. Smith, and S. Kambhampati, “Explicability? legibility? predictability? transparency? privacy? security? the emerging landscape of interpretable agent behavior,” in *Proceedings of the international conference on automated planning and scheduling*, vol. 29, 2019, pp. 86–96.
- [8] F. Li, D. C. Hogg, and A. G. Cohn, “Advancing spatial reasoning in large language models: An in-depth evaluation and enhancement using the stepgame benchmark,” in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 38, no. 17, 2024, pp. 18 500–18 507.
- [9] J. W. Strachan, D. Albergo, G. Borghini, O. Pansardi, E. Scaliti, S. Gupta, K. Saxena, A. Rufo, S. Panzeri, G. Manzi, *et al.*, “Testing theory of mind in large language models and humans,” *Nature Human Behaviour*, vol. 8, no. 7, pp. 1285–1295, 2024.
- [10] A. Lovett and K. Forbus, “Modeling visual problem solving as analogical reasoning,” *Psychological review*, vol. 124, no. 1, p. 60, 2017.
- [11] I. Rabkina, C. McFate, K. D. Forbus, and C. Hoyos, “Towards a computational analogical theory of mind,” in *Proceedings of the Annual Meeting of the Cognitive Science Society*, vol. 39, 2017.
- [12] C. Baker, R. Saxe, and J. Tenenbaum, “Bayesian theory of mind: Modeling joint belief-desire attribution,” in *Proceedings of the annual meeting of the cognitive science society*, vol. 33, no. 33, 2011.
- [13] M. Shum, M. Kleiman-Weiner, M. L. Littman, and J. B. Tenenbaum, “Theory of minds: Understanding behavior in groups through inverse planning,” in *Proceedings of the AAAI conference on artificial intelligence*, vol. 33, no. 01, 2019, pp. 6163–6170.
- [14] I. Rabkina and K. D. Forbus, “Analogical reasoning for intent recognition and action prediction in multi-agent systems,” in *Proceedings of the Seventh Annual Conference on Advances in Cognitive Systems*. Cognitive Systems Foundation Cambridge, 2019, pp. 504–517.
- [15] N. Rabinowitz, F. Perbet, F. Song, C. Zhang, S. A. Eslami, and M. Botvinick, “Machine theory of mind,” in *International conference on machine learning*. PMLR, 2018, pp. 4218–4227.
- [16] T. N. Nguyen and C. Gonzalez, “Cognitive machine theory of mind,” in *CogSci*, 2020.
- [17] Y. Mao, S. Liu, Q. Ni, X. Lin, and L. He, “A review on machine theory of mind,” *IEEE Transactions on Computational Social Systems*, 2024.
- [18] M. Verma, S. Bhambri, and S. Kambhampati, “Theory of mind abilities of large language models in human-robot interaction: An illusion?” in *Companion of the 2024 ACM/IEEE International Conference on Human-Robot Interaction*, 2024, pp. 36–45.
- [19] L. M. Hiatt, A. M. Harrison, and J. G. Trafton, “Accommodating human variability in human-robot teams through theory of mind,” in *IJCAI Proceedings-International Joint Conference on Artificial Intelligence*, vol. 22, no. 3. Barcelona, 2011, p. 2066.
- [20] A. Nematzadeh, K. Burns, E. Grant, A. Gopnik, and T. L. Griffiths, “Evaluating theory of mind in question answering,” *arXiv preprint arXiv:1808.09352*, 2018.
- [21] M. Le, Y.-L. Boureau, and M. Nickel, “Revisiting the evaluation of theory of mind through question answering,” in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 2019, pp. 5872–5877.
- [22] I. Rabkina, P. Kantharaju, M. Roberts, J. Wilson, K. Forbus, and L. Hiatt, “Recognizing the goals of uninspectable agents,” *Advances in Cognitive Systems*, 2020.
- [23] G. Bohning and J. K. Althouse, “Using tangrams to teach geometry to young children,” *Early childhood education journal*, vol. 24, pp. 239–242, 1997.
- [24] D. H. Clements and J. Sarama, “Early childhood teacher education: The case of geometry,” *Journal of mathematics teacher education*, vol. 14, pp. 133–148, 2011.
- [25] B. N. Verdine, C. M. Irwin, R. M. Golinkoff, and K. Hirsh-Pasek, “Contributions of executive function and spatial skills to preschool mathematics achievement,” *Journal of experimental child psychology*, vol. 126, pp. 37–51, 2014.
- [26] A. Ji, N. Kojima, N. Rush, A. Suhr, W. K. Vong, R. Hawkins, and Y. Artzi, “Abstract visual reasoning with tangram shapes,” in *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, 2022, pp. 582–601.
- [27] A. Langer, J. R. Wilson, L. Howard, and P. J. Marshall, “The influence of child characteristics on children’s behavior towards a social robot versus a human,” under review.
- [28] Y. Yang, A. Langer, L. Howard, P. J. Marshall, and J. R. Wilson, “Towards an ontology for generating behaviors for socially assistive robots helping young children,” in *Proc. of the AAAI Symposium Series*, vol. 2, no. 1, 2023, pp. 213–218.
- [29] D. A. Randell, Z. Cui, and A. G. Cohn, “A spatial logic based on regions and connection,” *KR*, vol. 92, pp. 165–176, 1992.
- [30] A. G. Cohn, B. Bennett, J. Gooday, and N. M. Gotts, “Qualitative spatial representation and reasoning with the region connection calculus,” *geoinformatica*, vol. 1, pp. 275–316, 1997.
- [31] J. R. Wilson, E. Krause, M. Scheutz, and M. Rivers, “Analogical generalization of actions from single exemplars in a robotic architecture,” in *Proceedings of the 2016 international conference on autonomous agents & multiagent systems*, 2016, pp. 1015–1023.
- [32] T. Fitzgerald, A. Goel, and A. Thomaz, “Human-guided object mapping for task transfer,” *ACM Transactions on Human-Robot Interaction (THRI)*, vol. 7, no. 2, pp. 1–24, 2018.
- [33] D. Gentner, “Structure-mapping: A theoretical framework for analogy,” *Cognitive science*, vol. 7, no. 2, pp. 155–170, 1983.
- [34] K. D. Forbus, R. W. Ferguson, A. Lovett, and D. Gentner, “Extending SME to handle large-scale cognitive modeling,” *Cognitive Science*, vol. 41, no. 5, pp. 1152–1201, 2017.
- [35] E. C. Grigore, A. Roncone, O. Mangin, and B. Scassellati, “Preference-based assistance prediction for human-robot collaboration tasks,” in *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2018, pp. 4441–4448.
- [36] J. R. Wilson, L. Tickle-Degnen, and M. Scheutz, “Challenges in designing a fully autonomous socially assistive robot for people with parkinson’s disease,” *ACM Transactions on Human-Robot Interaction (THRI)*, vol. 9, no. 3, pp. 1–31, 2020.
- [37] J. Fasola and M. J. Mataric, “Using socially assistive human-robot interaction to motivate physical exercise for older adults,” *Proceedings of the IEEE*, vol. 100, no. 8, pp. 2512–2526, 2012.