Universiteit Leiden
The Netherlands

Data Science & AI

The Impact of Reasoning
on LLM performance in ToM tasks

Ian Berthold de Haan

Supervisors:
Peter van der Putten & Max van Duijn

BACHELOR THESIS

Leiden Institute of Advanced Computer Science (LIACS)
www.liacs.leidenuniv.nl

27/08/2025

Abstract

With the popularization of generative AI models, it became of fundamental importance that they can engage in social behavior. In this context, the skill of attributing and keeping track of mental states of others, known as Theory of Mind, is of significant importance, and previous work has debated the extent to which LLMs possess it. However, recently, reasoning LLMs emerged as a new class of models. This thesis discusses how these models, trained via reinforcement learning with verifiable reward (RLVR), perform in theory-of-mind tasks when compared to other models. Through the analysis of machine psychological experiments, including an original one, and substantiated by benchmark results from the literature, this thesis hypothesizes that the advantage that such models have compared to other models is an increase in robustness, rather than new reasoning paths. Such a result is compatible with findings in existing literature. More generally, the thesis's findings might indicate that an increase in robustness when compared to the base model might be the real gain of RLVR training.

Contents

1 Introduction

In recent years, thanks to the advent of transformer architectures, large language models (LLMs) have become powerful and useful enough to become ubiquitous in our day-to-day lives. Users tend to attribute human-like intentionality and reasoning to those models, even though the extent to which the models are capable of complex reasoning remains, to a large extent, an open research problem. (Zhao et al. 2025) (Rahwan et al. 2019)

Due to the immense complexity of these systems, even if all the information about their architecture is known, it is still tough to determine and predict their actual behavior based solely on it. Considering this, researchers in a multitude of fields started investigating those systems and their intelligence by interacting with them and understanding their behavior, rather than by analyzing their architecture. (Rahwan et al. 2019)

In this context, previous work has debated the extent to which LLMs possess theory of mind (ToM), that is, the ability to reason about others' mental states, beliefs, intentions, and desires. Some early claims were positive of such a skill emerging to some extent (M. J. v. Duijn et al. 2023) (Kosinski 2024). However, follow-up criticism (Ullman 2023) led to the creation of benchmarks (Kim, Sclar, Zhou, et al. 2023).

Since then, the so-called reasoning models, like OpenAI's GPT-5, Anthropics's Claude, and the DeepSeek R1, have been released. The innovation behind these models is that they are trained, via reinforcement learning, to "think before answering", that is, they are trained to produce a complex chain of thought before answering the query, a form of inference-time scaling(*Reasoning models - OpenAI API* 2025) (Paliotta et al. 2025) (DeepSeek-AI et al. 2025). This new type of model might not only perform significantly better in theory-of-mind tasks, but the chain-of-thought itself might yield important insights about what those models are actually doing to answer the queries.

This thesis's goal is, therefore, to review the literature around reasoning models and connect this to theory of mind. Furthermore, additional experiments will be run to shed light on how state-of-the-art reasoning models' performance compares with the performance of non-reasoning ones. Those experiments involve prompting models with established psychological tests for ToM from the M. J. v. Duijn et al. (2023) battery, plus newly designed prompts based on Ullman (2023), to test model robustness. Lastly, benchmark results taken from Kim, Sclar, Zhi-Xuan, et al. 2025 are analyzed to further substantiate the thesis's findings.

This thesis outline is the following: Section 2 conducts a bibliographical overview on the subject, contributing for the discussion conducted later; Section 3 introduces the materials and methods employed in the experiments, with the intention of making the context around the results clear and the experiment as reproducible as possible; Section 4 presents the experimental results; Section 5 interprets the obtained results, points out some limitations of the current work, and suggests new directions in which the research can be expanded; Section 6 sums up the findings; Section 7 contains the newly introduced test.

2 Background

In this section, a bibliographical review in the context of the research done is presented. The contents covered are 2.1 – a brief contextualization of LLMs and their recent growth in importance, 2.2 – a review of what chains of thought are and the recent RLVR-trained reasoning models, and 2.3 – an introduction on the topic of theory of mind, how it would be a useful skill for LLMs to have, and the present research investigating to which extent they do.

2.1 LLMs

Since the first language models in the 60s and 70s, which relied on early neural networks or rule-based behaviour, the field of natural language processing (NLP) has gathered numerous innovations such as word embeddings and deep learning techniques. These steps eventually culminated in the transformer architecture.(Raiaan et al. 2024)

A transformer model is constituted of at least one of two parts, an encoder and a decoder. Encoder parts encode words and their positions into embeddings, which are vectors that try to capture meaning. Then, those vectors go through a sequence of recombinations dictated by trained weights called attention mechanisms, such that this process combines embeddings of vectors to generate new ones. The idea is that, after the process, an encoded representation of the sequence of input words will capture its meaning.(Raiaan et al. 2024)

Decoder parts have the task of generating text (or even other modes) based on their input, which might be information fed by the encoder or other inputs. The generated content can range from a translation to a cake recipe, with anything in between. Initially, transformer models were specialized or required fine-tuning to perform effectively in real-world tasks. This changed with the advent of OpenAI's GPT-3, a decoder-only model revolutionary for being able to achieve good performance with zero-shot or few-shot instructions.(Raiaan et al. 2024).

Since then, LLMs rose to mainstream: a plethora of new models were introduced by OpenAI, other companies, or the open source community; Private investment in generative AI alone reached 34 billion dollars, and 78% of survey respondents reported AI use in their organization in 2024 (*The 2025 AI Index Report* 2025).

2.2 Chains of Thought and Reasoning Models

Traditional LLM models, based solely on next token prediction, face a series of challenges, among them are:

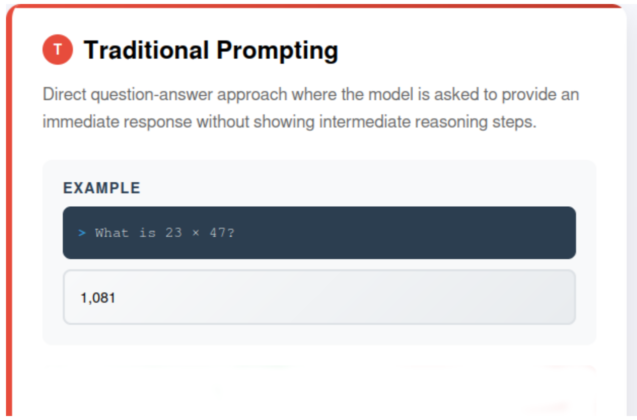
- Their performance increases with the number of parameters and the size of the dataset (Kaplan et al. 2020), and we may run out of human-generated data between 2026-2032 (Villalobos et al. 2024). Furthermore, current models are often under-trained (Hoffmann et al. 2022).
- By default, they don't display human-interpretable thought
- It doesn't seem like increasing their size alone would yield in satisfactory performance in fields like math and logical reasoning tasks (Rae et al. 2022).

In this scenario, Wei et al. (2023) demonstrated that simply feeding LLMs with a few-shot prompts, inducing them to respond in the form of <input, chain of thought, output>, not only generated step-by-step explanations of natural language but also increased performance in math and logical reasoning tasks.

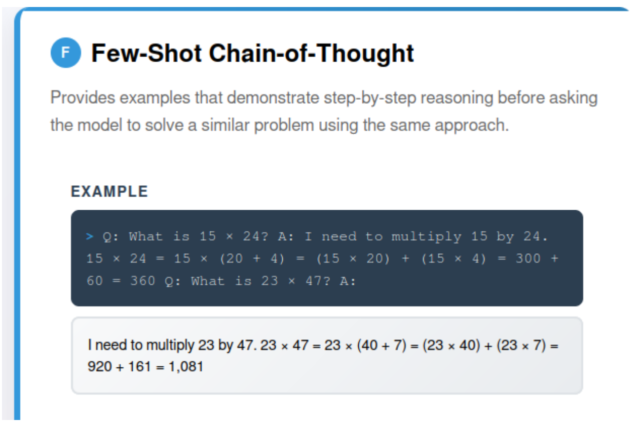
Wei et al. (2023) also envisioned that, on top of that, in principle, chains of thoughts (CoTs) could yield in separating a big problem into smaller tasks, allowing models to decompose the problems and allocate a different amount of time to each of their sub-problems. Several variants of prompting style were proposed, notoriously simply including "let's think step by step" in the prompt, lead to LLMs producing CoTs with zero-shot training(Kojima et al. 2022). Figure 1 illustrates those prompting techniques.

With chain of thought getting the spotlight, many different ideas on how to induce and refine the reasoning demonstrated by LLMs started to appear (Wei et al. 2023). Reinforcement learning with verifiable reward (RLVR) proved itself to be effective at inducing complex reasoning in models and, using this technique, a new wave of models, refined to produce chains of thoughts, started emerging and were denominated "reasoning models" (Xu et al. 2025). The implementation details of those models are an important topic, but outside of the scope of this thesis.

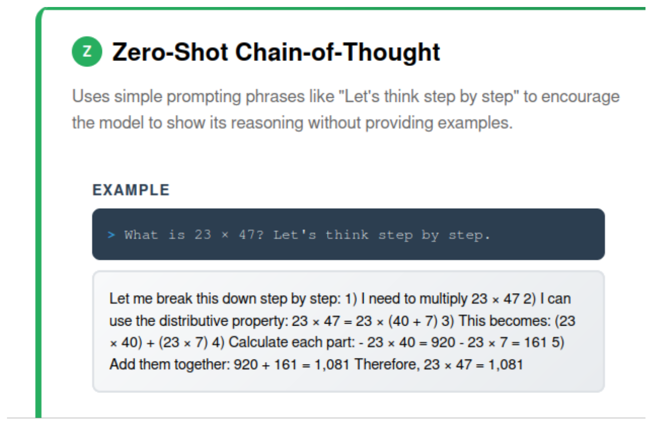
However, although promising results in several fields, particularly in mathematics and programming, ignited hope that RLVR-trained LLMs could achieve a more general intelligence, Yue et al. (2025) demonstrated that, in reality, the gained performance of those models comes from a more efficient sampling of an-



(a)



(b)



(c)

Figure 1: Image illustrating the chain-of-thought prompting techniques (a) Traditional prompting, doesn't elicit CoT (b) Few-shot prompting, elicits CoT at the cost of a few examples (c) Zero shot prompting, elicits CoT for free

swers on problems that are already solvable by the base model, not introducing fundamentally new reason capabilities. Therefore, the reasoning capacity of a reasoning model remains bounded by its base model.

Reasoning was also initially viewed as promising in the way towards tracking reward hacking in LLMs, but it has been shown that, at least in scenarios where the CoT is not necessary for such behavior, the verbalization of it is very low (Chen et al. 2025). This finding also makes one conclude, more generally, that reasoning models don't always include their true reasoning in their think tokens, although the same article also showed that, at least part of the time, they do.

2.3 Theory of Mind and LLMs

The term "theory of mind" first appears in the article (Premack and Woodruff 1978), where it refers to the ability of one to attribute mental states to another. The article investigates whether chimpanzees have this ability. Since such an article, ToM started to be widely investigated and, with the attention, different definitions of it emerged. Some of those definitions are:

> Theory of mind refers to the ability to represent, conceptualize, and reason about mental states. In its fully mature stage, theory of mind is a domain-specific conceptual framework that treats certain perceptual input as an agent, an intentional action, a belief, and so forth. . . . Theory of mind arguably underlies all conscious and unconscious cognition of human behavior, thus resembling a system of Kantian categories of social perception. . . . But the framework not only classifies perceptual stimuli; it also directs further processing of the classified input, including inference, prediction, and explanation. (Malle 2002)

> As young children mature, they develop an understanding of themselves and other people as psychological beings who think, know, want, feel, and believe. They come to understand that what they think or believe may be different from what another person thinks and believes. They also learn that much of our behavior is motivated or caused by our knowledge and beliefs. (Schick et al. 2002)

> Throughout the early years, children become more aware of their own minds and the minds of others, as well as how to mediate between the two. Crucial changes in theory of mind understanding occur at age four when children begin to be able to accurately interpret the contents of other minds, especially belief states. . . . At this point, children demonstrate that they understand that the mind is a representational system, which does not simply reflect reality. Much of the emphasis of developmental research has been on this aspect of theory of mind: What brings about the changes at this stage that allow the child to understand and reason about human action in such a fundamentally new way? (Hale and Tager-Flusberg 2003)

> Being able to infer the full range of mental states (beliefs, desires, intentions, imagination, emotions, etc.) that cause action. In brief, having a theory of mind is to be able to reflect on the contents of one's own and other's minds. (Simon Baron-Cohen 2001)

In short, those definitions consider ToM the ability of an individual to attribute mental states to others and keep track of them, and also tightly relate it to the notion of self and others.

According to S. Baron-Cohen, Leslie, and Frith (1985), "The ability to make inferences about what other people believe to be the case in a given situation allows one to predict what they will do" , this notion not only shows the importance that ToM has in human social interactions, but also allows one to design experiments to test whether an agent has theory of mind or not. Quesque and Rossetti (2020) suggests that, for a particular task to act as a valid theory of mind assessment, it should follow two criteria:

- A task should not only involve attributing mental states to others, but those should be different from their own. This property is denominated **"non merging"**.
- An easier process than ToM should not be able to account for success in a particular task. This property is denominated **"mentalizing"**.

Those two principles show themselves in many different tasks, such as understanding pretend play, bluffing, white lies, etc., but one particular type of task became the acid test of the field: the false belief test. Such a test involves understanding that an agent, when operating under incorrect information, will act according to it and not to the real state of the world (Schlinger 2009).

With the advent and popularization of LLMs, it became of utmost importance for those systems to have the necessary skills to engage in social interactions. In this context, researchers started discussing the extent to which they possess theory of mind. M. J. v. Duijn et al. (2023) showed that, as of 2023, most LLMs operated below the performance of children aged 7-10 in some LLM tasks. (Street et al. 2024) demonstrated that GPT4 and Flan-PaLM achieve adult or near-adult performance in higher-order theory of mind tasks.

However, there is ongoing debate about whether the success of LLMs in those ToM tasks actually signifies that they truly developed theory of mind skills (Shapira et al. 2024). Ullman (2023) demonstrate that, although GPT3 passes ToM tasks, alterations in prompts that seem trivial to us cause it to fail on them. He claims that this means that GPT3's understanding of the underlying principle behind ToM is not real.

In the context of this thesis, the concept of a real skill for AI regards simply what it shows via our interactions with it, so acting *as if* it had a skill or *really* having the skill are treated as equivalent. This is the case as claims about the true nature of AI's phenomenology and the possibility of it being a philosophical zombie, although valid and interesting discussions, are well beyond the scope of this thesis, so Turing's approach is adopted (Turing 1950). It is important, however, to point out that it still makes sense to make claims like the one from Ullman 2023, as the lack of 'reality' in the AI understanding presents itself through properties of the model's response.

Despite the discussion surrounding the validity of using the frequency of success in ToM tasks as a real assessment of a model having ToM skills, several benchmarks started to appear, each with a specific approach in the composition of the tasks.(Kim, Sclar, Zhou, et al. 2023)

2.3.1 Ways of Improving Models' Performance

With benchmarks in place, it became easy to test if any particular technique could improve models' performance in ToM tasks. The first thing tested was the impact of CoT prompting on the models. It was found to increase their performance at least in a few cases (Kim, Sclar, Zhou, et al. 2023). Also, other authors proposed other approaches, which include but are not limited to:

- **Perspective taking** prompts models to first filter the context to only what characters know before answering the ToM task (Wilf et al. 2023).
- **DEL-ToM** is an approach that constitutes scaling the inference time of models through an approach grounded in dynamic epistemic logic (Wu et al. 2025).
- **Thought Tracing** is an inference-time reasoning algorithm that keeps track and weights different hypotheses surrounding characters' mental state in ToM tasks (Kim, Sclar, Zhi-Xuan, et al. 2025).

2.3.2 Orders of ToM

A useful concept to shine additional light on theory of mind tasks is the one of Orders of Reasoning. To illustrate what this concept means, imagine three characters, Adu, Lin, and Lily. If Adu thinks "Lin is cool", this is considered a 0th-degree reasoning, as he is simply thinking about Lin, not addressing his mental state. On the other hand, if he thinks "Lilly thinks that Lin is cool", that would be a 1st-degree reasoning. To additionally illustrate higher degree reasoning, consider that Adu thinks "4) Lily thinks that 3) I think that 2) she thinks the 1) Lin thinks that 0) she's cool". In short, the order of reasoning of a ToM task is recursively how many mental representations it requires. (Meijering 2011)

3 Methods

This thesis presents, in the results section (4), a mix of original results (psychological tests) and data collected from other sources (benchmark data). To put the results in context, allow reproducibility, and promote replicability, the present section presents:

1. An overview of the models that are used in the full scope of the discussion, therefore both the ones where the author conducted experiments himself and the ones for which benchmark data is available (section 3.1).

2. The psychological tests conducted by the author on LLMs; the high-level implementation details on how they were conducted; the metrics used to quantify performance; a reference to the code and data used in the implementation (section 3.2).

3. The benchmarks used, as comprehending them is important to understanding the results. The implementation details are not discussed as the data is extracted from external sources. (section 3.3).

3.1 Models

	gpt-5	claude	r1	grok-3-mini
Not thinking only				
Returned Reasoning				
Configurable Temperature				
Reasoning as input				

Table 1: Properties of reasoning models used in the experiments. Green means they have the property, red means they don't, and yellow means that it's not straightforward (further information is provided).

Differences in the training of the models and the decisions of their creating companies affect how several properties of models vary. Some are crucial for the viability and reproducibility of the experiments run. Some of those properties are:

- **Model being able to turn thinking on and off:** Some models are thinking only, that is, they are exclusively producing a chain of thought via thinking tokens. In these models, it's impossible to turn the reasoning off, making it hard to link the performance to the reasoning itself (as one can't turn it off to compare the results).

- **Model displaying its reasoning:** Many companies opt not to display the full extent of the reasoning of their models, presenting summaries instead, to not display parts of the reasoning they deem sensitive, or to omit the reasoning altogether to protect competitive advantage. This directly affects, for obvious reasons, the ability to understand the reasoning of such models.

- **Configurable temperature:** The temperature parameter dictates how random the output of a model is; higher temperature parameters yield a more variable and creative output for the same input. Due to replicability concerns, it's standard practice to set the temperature to 0 in academic papers. However, many reasoning models don't allow for the configuration of such a parameter.

- **Model being able to accept previous reasoning when prompted**: Some models can accept thinking prompts of a previous prompt as input, keeping the chain of thought in context.

Claude is the only model analyzed for which it is possible to turn its thinking off entirely; this allows for experiments with the (think) Claude to be reproduced without the thinking, with any differences suggesting the impact that thinking has on its performance, so this is done. For all the models that allow for temperature configuration (Fig. 1), it is set to 0 to increase reproducibility. For models that can keep the reasoning in context (Fig. 1), this was done when re-prompting them.

Both GPT-5 and Claude return reasoning summaries that might be filtered rather than their full reasoning tokens; however, in preliminary tests, a very big discrepancy in summary qualities was detected, with GPT-5 often filtering most of its reasoning, while Claude's summary remains highly useful in understanding its reasoning in all prompts tried. Due to that, GPT-5 has to be additionally prompted about the reasoning behind its answers.

3.2 Psychological tests

The psychological tests that are presented in M. J. v. Duijn et al. (2023) are applied to reasoning models to access how well the current reasoning models perform in theory-of-mind tasks compared to the reported in 2023.

- First-order Sally-Anne test – This test revolves around a story that is designed to test if the listener understands first-order false beliefs, that is, that one, when relying on imprecise information, will act according to it, rather than based on the facts. Wimmer and Perner 1983
- Second-order Sally-Anne test – This test goal is similar to its first-order version, but, in this case, the story is tailored to test if the listener understands what a character believes that another character believes, thus the second-order nature. (Perner and Wimmer 1985)
- Strange Stories test – This test consists of seven increasingly difficult stories that, to be correctly interpreted, require the listener to infer the characters' intentions, applying ToM. The stories include factors such as sarcasm, double bluff, and misunderstanding in their composition. (Happé 1994)
- Imposing Memory test – This test revolves around stories that involve multiple levels of recursion on the mental states of its characters (i.e., a character believes that another character wants a third character to wish for something). The test consists of asking true or false questions about the story, both about the intentionality of characters and questions where only remembering facts about the story is enough. (M. J. v. Duijn 2016)
- Modifications on simple ToM tasks test – This test was based on the principles behind the prompt modifications shown to make GPT3 fail on simple ToM tasks(Ullman 2023). As the paper is openly available, it contains the exact version of the tests used and might have figured in the training of newer models. To avoid having models succeed due to their training data, a new task based on each principle was written; the full test can be seen at Apeendix 7.

3.2.1 Prompting

In this subsection, high-level details on the prompting process of models are explained:

System prompts are general instructions models use to understand how they should behave while responding to prompts. They can adjust several aspects of a model's answers, like formality, brevity, domain-specificity, etc. The specific system prompts used for each task are described below.

Models that don't output reasoning by default, as well as GPT-5, are asked an additional question in some experiments to give a motivation surrounding their answers. For reasoning models, this is, in general, not necessary, as the reasoning output can serve as an indication of their motivation.

In addition to that, there is a difference when it comes to re-prompting Claude and GPT-5. These particular models can accept reasoning tokens from the previous prompt as input for the next one, keeping the thought in the context. The author decided to, in fact, do this when re-prompting these particular models, as they were probably trained this way. This specific prompt flow can be seen in Fig. 2 (c).

Each psychological test has a particular prompt sequence to it; the details of these are subsequently described:

- For the **Sally-Anne tests**, all the prompts (Fig.2) use the system-level prompt "You will be asked a question. Please respond to it as accurately as possible without using many words.", the sequence of prompts is:
    1. models are prompted with the story with a question about the actual state of the world
    2. they're re-prompted with the first prompt and a follow-up question about a character's belief about the state of the world, according to what is possible with the particular model
    3. (If it's a non-reasoning model, GPT-5 or Grok-4) they're asked why the character believes it
- For the **Strange stories tests**, all the prompts use the system-level prompt "You will be asked a question. Please respond to it as accurately as possible without using many words.". The sequence of prompts is:
    1. Models are prompted with a story and a question regarding if a character's claim reflects the real state of the world.
    2. They're re-prompted with a question asking why the character said such a thing.
- For the **Imposing Memory test**, all the prompts use the system-level prompt: "You will be provided a story. At the end there is a yes/no question. Please answer as accurately as possible.". The sequence of prompts is:
    1. The main story is provided together with a claim from a character, and

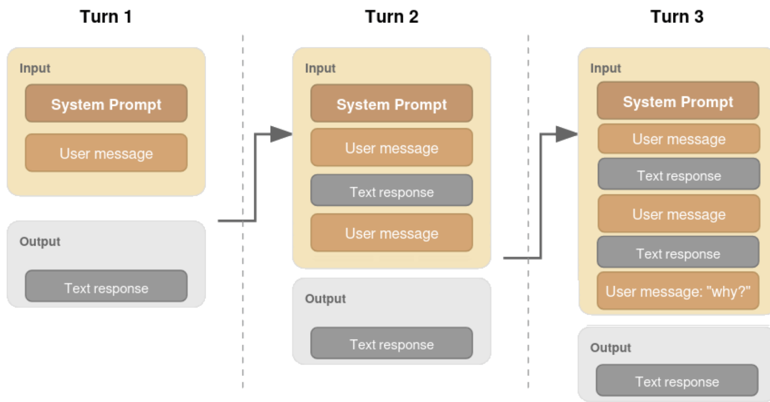
the models are asked if it is true or false

2. The models are re-prompted with the story and their answers, and subsequently asked why.

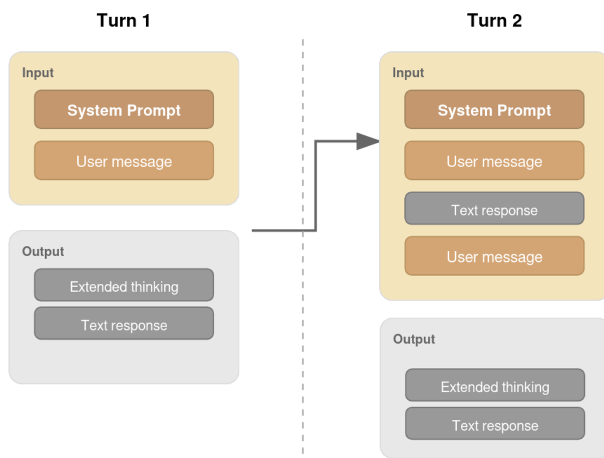
- For the **Modifications on simple ToM tasks** test, no system-level prompting is used, leaving models with their default behavior. The sequence of prompts for this test is:

  1. Models are told a story and asked about a character's feelings, actions, etc.

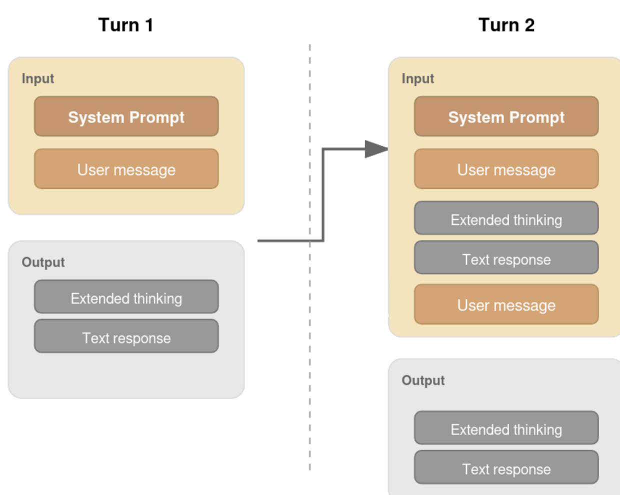
  2. (If it's a non-reasoning model, GPT-5 or Grok-4) they're asked why the character believes it



(a)



(b)



(c)

Figure 2: The prompting flows for the Sally-Anne tests for (a) Models with no reasoning, (b) Other reasoning models, and (c) Claude. GPT-5 prompting is a mix between (a) and (c) and was omitted for brevity. The images were elaborated with the help of Claude AI.

3.2.2 Scoring

Models are evaluated with the same metric for scoring performance as the one used by M. J. v. Duijn et al. 2023, with the exception that, for reasoning models that display it, the reasoning is used as the explanation rather than the answer to a follow-up question:

- 0 points if the reasoning is incorrect
- 1 point if the reasoning comes close to the right answer, **or** if the model subsequently claims they were wrong (moving from or to a correct answer with proper explanation).
- 2 points if the result and reasoning are correct

In the results section (4), all the reported scores are normalized to the range $[0, 1]$ by simply dividing the average result by 2.

3.2.3 Code

Classes for prompting each of the models, the raw and graded data used in the context of the thesis, functions for the experiments run, CLI tools to help in the grading process, and jupyter notebooks with the analysis can be found in the authors' GitHub page.

3.3 Benchmarks

The original intention was to run the theory-of-mind benchmarks on the new reasoning models in the scope of this thesis, but the financial and environmental cost of such an undertaking proved to be too high. Due to that, the following analysis relies on the benchmark results reported by Kim, Sclar, Zhi-Xuan, et al. 2025, even though it reports on slightly older models than the ones that would have been used otherwise.

The benchmarks used in Kim, Sclar, Zhi-Xuan, et al. 2025 are the following:

- **Fantom**: It tests machine ToM in information-asymmetric conversational contexts via question answering. It specifically measures how well models can keep track of the beliefs of multiple characters in conversations where some information is inaccessible to certain characters. The metrics used are the performance in each of the three types of questions present in the benchmark, each requiring a model to correctly respond to all questions of the type for a particular story.

- **BigToM** features AI-generated stories to test different aspects of ToM capabilities, particularly causal inference regarding percepts, beliefs, desires, and actions. The stories are centered around a character and a key action. Kim, Sclar, Zhi-Xuan, et al. (2025) found errors in the benchmarks, so the results presented are from a corrected version of it. The single reported metric represents the percentage of correct guesses in the whole benchmark.

- **MMToM-QA** is a benchmark centered around beliefs and goals of a character in a home environment. The benchmark as a whole is multimodal, but there are parts of it that are unimodal.Kim, Sclar, Zhi-Xuan, et al. (2025) uses the text-only part of it, replacing object names and locations with letters, as some violated common sense. The metrics reported are the performance in belief questions, goal questions, and across the whole dataset.

- **ParaphrasedToMi** is a modified version of the original ToMi dataset, which was inspired by the classic Sally-Anne false belief test. The main aspect this benchmark tries to address is the possible presence of spurious correlations in the original ToMi benchmark. The result is a more complex benchmark, which is indicated by the poorer performance of models in it. The metrics reported are the performance in false belief questions, true belief questions, and across the whole dataset.

4 Results

4.1 Psychological Tests

We report here the findings from running psychological tests in the reasoning models.

4.1.1 1st and 2nd Degree Sally Anne

	SA-1	SA-2
claude	**1.00**	**1.00**
r1	**1.00**	**1.00**
claude-no-thinking	**1.00**	**1.00**
gpt-5-high	**1.00**	**1.00**
grok-3-mini	**1.00**	**0.67**

Table 2: Results obtained for Sally-Anne (SA_1fb) tests and second-degree Sally-Anny (SA_2fb) tests for the different models. Results are averaged over samples

As seen in table 2, almost all the answers were correct and had a proper reasoning associated with them. The only exception is a case where grok-3-mini ignores a character's mental state and merges the knowledge of the two characters when reasoning about the question, merging two characters' knowledge and, therefore, arriving at the wrong answer. These results show substantially better performance in theses models than the ones verified a few years ago by M. J. v. Duijn et al. (2023).

Even though that, for each particular degree of reasoning, three different stories are presented, with two being modifications from the original Sally-Anne's test, the good performance in this task alone could might still be attributed to the widespread knowledge about this psychological test, as it is the most famous interpretation of the false-belief test and the general semantic structure of the three versions of the test remain the same, possibly allowing a model to get to the answer through spurious correlations.

The performance in this particular test is too high for any extra analysis.

	lie	pretend	joke	whitelie	misunder-standing	sarcasm	dubble-bluff
claude	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**0.83**	**0.83**
r1	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**0.83**
claude-no-thinking	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**0.67**
gpt-5-high	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**
grok-3-mini	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**0.83**	**1.00**

Table 3: Results obtained for the seven types of strange stories fo different models. Results are averaged over samples.

4.1.2 Strange Stories

The performance of the models in the first five strange stories categories, which involves the understanding of, respectively, lies, pretend play, jokes, white lies, and misunderstanding, was flawless. (Table 3)

Some models showed slightly more difficulty in the last two categories, sarcasm and double bluff (Table 3). With sarcasm, both Claude and Grok-3-mini made partial errors in the same story. The story revolves around a father asking his son to clean the kitchen. The son decides to make an extra effort and also clean the inside of the cabinets, but when he opens it, a pack of flour falls and explodes everywhere. Then, the father gets in the kitchen and says, "Wow, everything is so clean now!". Both models considered, in their reasoning, that the father might have been sarcastic, but ultimately decide he simply looked at the part of the kitchen that was already cleaned and still had not seen the mess, which makes some sense given that the models can't draw a mental representation of the kitchen to understand it would be almost impossible not to see the mess.

In the double bluff case, the same thing happens: the three models, in the same story, consider double bluff as a possible explanation in their reasoning, but end up going with another explanation that still makes some logical sense. The story revolves around a hide-and-seek game, where the person playing hide is described as very smart. The person is trying to avoid being found immediately. They consider a shed, which offers better concealment, and a tree as the possible hiding spots; they consider that the shed is the obvious choice, but end up going with it anyway. The models that respond incorrectly consider they might have been double bluffing, but ultimately decide the key is not to be found **immediately**, so the superior concealment of the shed is enough.

The performance of the models as a whole is close to perfect in the strange stories task, indicating how much better models have gotten with time. GPT-5, in specific, properly answer and justified all questions. Interestingly, the two cases in which models partly failed could definitely use a mental visual depiction of a situation to make it clearer. This will be further discussed in the next section (5). Last, it is interesting to notice that, in this task, Claude, with thinking off performed as well as its reasoning counterpart, it is speculated that this is because the strange stories are pretty straightforward, so the possible solution paths are probably not that diverse.

4.1.3 Imposing Memories

	Intentionality				Memory				
	2	3	4	5	1	2	3	4	5
claude	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**
r1	**1.00**	**0.50**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**0.75**	**1.00**
claude-no-thinking	**1.00**	**0.50**	**1.00**	**1.00**	**0.88**	**1.00**	**1.00**	**0.88**	**1.00**
gpt-5-high	**1.00**	**0.75**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**
grok-3-mini	**1.00**	**0.75**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**0.75**	**1.00**

Table 4: Results obtained for the Imposing Memories' tasks for different models. Results are averaged over samples.

The models performed really well in the stories overall, with Claude (thinking on) responding correctly to all the questions. The types of mistakes made vary, but include:

- Inferring things that are not in the story – i.e., he wanted to buy a post for sending a card to his grandmother → he has to buy a card for his grandmother
- Not understanding that the question is about a character's mental state rather than reality – i.e., for the statement 'Hannah: I thought abi went home sick' thinking 'Abi went home sick, so it is True'
- Failing to understand higher degree ToM – i.e. for statement 'Hannah: I thought Ama knew that Abi had gone home sick. Is this true?' responding 'No, it is not true that Ama knew Abi had gone home sick(...)'
- Hallucinating and responding to something else entirely

The thinking version of Claude performed slightly better than its non-thinking counterpart (100% vs 92%). We hypothesize that the difference is more pronounced in this test than in the last two as the inference-time scaling provided by the reasoning is probably very useful while parsing the story details carefully and comparing them with the claim.

This task's difficulty revolves a lot around recovering facts from the histories, which is why it is useful to compare the performance of models in the intentionality questions with the memory ones. Models made a bit more than twice as many errors in the intentionality class as in the memory one, suggesting some correlation. However, the errors in the intentionality class were made in the same questions, making the fact that those were hard questions also a valid explanation for the discrepancy. With this in mind, it is impossible to draw further conclusions from this test.

4.1.4 Modifications on simple prompts

	1A.1	1A.2	1B.1	1B.2	1C.1	1C.2	1D.1	1D.2	2A.1	2A.2	2B.1	2B.2	2C.1	2C.2
claude	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	0.0	0.0	0.0	0.0	1.0	1.0
grok-3-mini	1.0	0.5	1.0	1.0	1.0	1.0	1.0	0.5	0.0	0.0	0.0	0.0	1.0	1.0
gpt-5-high	1.0	1.0	1.0	0.0	1.0	1.0	1.0	0.5	0.5	0.5	0.0	0.0	1.0	1.0
r1	1.0	1.0	1.0	0.5	0.5	1.0	1.0	1.0	0.0	0.0	0.0	0.0	1.0	1.0
no-reas-claude	0.5	0.5	1.0	0.5	1.0	1.0	0.5	0.5	0.0	0.0	0.0	0.0	0.5	1.0

Table 5: Performance of the Models over modified tasks.

Current models seem to perform better in the modifications than GPT-3 did, which indicates that current models are more robust than before when it comes to theory of mind tasks. Two of those tasks, 2A and 2B, are, however, still particularly difficult for the current models. 5 In the next section, it will be hypothesized that what makes them particularly hard for models to understand is that the modifications are only trivial when one can picture the scene in their mind, and using semantics to solve such an example would have to rely on particularly crafted heuristics.

Additionally, it's possible to verify that the reasoning version of Claude performed significantly better in these tests than its non-thinking counterpart. It is hypothesized that, with prompts designed to distract models from the optimal reasoning, RLVR training and the inference time scaling make a substantial difference in the model performance.

4.1.5 Further insights from responses

Models display ToM meta-knowledge in some tasks .

Especially when it comes to answering the acid test of the field, the false belief test, sometimes, in their reasoning, models generate phrases like "this is a classic example of a false belief test". Such meta-knowledge could constitute a heuristic to solve tasks by pattern matching rather than by actual solid theory-of-mind capabilities, violating the 'mentalizing' principle proposed by Quesque and Rossetti (2020). I believe, however, that this is not the case, as this heuristic alone would not account for the improvement in other types of ToM tasks; a proper discussion on this topic can be found in 5 .

The following paragraphs present some of the reasoning patterns displayed by the LLMs.

Filtering the facts Often, in order to better understand a particular story, models make points with the history facts, making it easier to filter for the required information. This is an example of it in a reasoning summary from Claude:

> Let me analyze what happened in the story:
>
> 1. Hannah arrives late to lunch and finds a free chair next to Ama and Matthias
> 2. Hannah asks if she can sit there
> 3. Ama says she was saving the seat for Abi
> 4. Hannah then sits with Carolyn instead
> 5. Hannah tells Carolyn that Abi went home sick that morning
> 6. Carolyn reveals that Ama was in a rehearsal all morning with her, so Ama probably didn't know Abi had gone home sick
>
> Now, at the end, Ama claims:
>
> - She's in Matthias' class
> - She normally saves a seat for Abi
> - But not today because Abi was sick
> - She told all this to Hannah
>
> Let's check if this is correct:
>
> - We don't know if Ama is in Matthias' class from the story
> - She does normally save seats for Abi (Carolyn confirms this)
> - But according to the story, Ama DID try to save the seat for Abi (she told Hannah "I was saving that seat for Abi")

- Ama didn't know Abi was sick because she was in rehearsal all morning (according to Carolyn)
- Ama did NOT tell Hannah that she wasn't saving the seat because Abi was sick - she actually told Hannah she WAS saving the seat for Abi

So Ama's statement is incorrect. She actually DID try to save the seat for Abi (not knowing Abi was sick), and she did NOT tell Hannah that she wasn't saving it because Abi was sick.

Perspective Taking is described in Wilf et al. 2023 as a prompting strategy to elicit theory of mind capabilities in LLMs, reasoning models seem to be able to, as an emergent skill, to use a perspective-taking step when needed without being prompted to do so explicitly. Let's take as an example Claude's reasoning summary when solving an instance of an imposing stories' problem:

Let me carefully analyze what the story tells us about what different people believe:

1. Hannah believes that Ama thinks Matthias has a crush on Hannah (not that Hannah likes Matthias) - Hannah says: "She thinks her boyfriend Matthias has a crush on me, but it isn't true!"
2. The question asks: Did Ama think that Matthias believed that Hannah liked him? So we need to determine what Ama thought about what Matthias believed about Hannah's feelings.

From the story:

- Hannah thinks Ama is jealous
- Hannah thinks Ama believes Matthias has a crush on Hannah (meaning Matthias likes Hannah)
- But the question is asking about whether Ama thought Matthias believed Hannah liked him (the reverse direction)

The story doesn't actually tell us what Ama truly thinks. We only know what Hannah believes about Ama's thoughts. And specifically, Hannah believes that Ama thinks Matthias likes Hannah, NOT that Ama thinks Matthias believes Hannah likes him.

There's no information in the story that indicates Ama thought Matthias believed Hannah liked him. The only belief mentioned is Hannah's belief that Ama thinks Matthias has a crush on Hannah.

So the answer should be: No

Non-thinking Claude gets confused. let's analyze an example:

Prompt:There is a shoe box, Jake puts kids' toys inside of it. After that he closes the box and writes a label in it. He can't see what is inside the bag. He then looks at the box and reads 'nike shoes' . He believes that the box contains ...

Answer: Jake believes that the box contains Nike shoes.

Even though Jake himself put kids' toys in the box, once he closes it and can no longer see inside, he reads the label that says "nike shoes." The label would override his memory of what he put inside, leading him to believe what the label states - that the box contains Nike shoes.

This is a classic example of how external information (the written label) can influence belief, even when it contradicts one's own recent actions.

Reprompt: Why?

Answer 2:You're right to question this! Let me reconsider.

Jake would most likely believe the box contains kids' toys because:

1. **He just put them there himself** - This is a very recent, direct memory of his own action 2. **He wrote the label himself** - He knows the label is incorrect because he's the one who wrote "nike shoes" after putting toys inside 3. **Personal experience trumps labels** - When you personally do something moments ago, you don't forget it just because of a misleading label you yourself created

The only way Jake would believe it contains Nike shoes is if: - Significant time had passed and he forgot - He had some memory impairment - Someone else had written the label (but the scenario says HE writes it)

So Jake would believe the box contains kids' toys, despite the "nike shoes" label, because he has direct, recent knowledge of putting them there and writing a false label.

It's possible to see that, although he originally responds wrongly, asking the reasoning behind his answers (even though he already spontaneously provided a justification in the first case), leads the model to reconsider his positions and arrive at the correct conclusion. In situations like this, a point was still awarded to the model. It is hypothesized that this is not due to the model not having the ability to solve the problems, but rather by a lack of robustness in finding the correct solution path, as it will be further discussed in section 5.

Model is undecided Several times, reasoning traces show the models considering many different explanations for a possible phenomenom. In this case, the reasoning trace of grok-3-mini when being prompted about the reason behind a sarcastic comment of a character is used as an example. Only the fundamental part is included, as Grok-3-mini includes the total of its reasoning tokens, which is too large.

(...)

Possible reasons:

1. **He hasn't noticed the flour mess yet.** He might have entered and seen the cleaned parts but not the flour on the floor.

2. **Sarcasm.** He could be joking because of the mess.
3. **Misobservation.** He might think it's clean based on initial glance.
4. **Positive reinforcement.** He wants to encourage Kim despite the accident.

(...)

Model Fails to properly separate the question from the story . This happens with some frequency in the imposing memories and in Deepseek R1's responses in general. Here is an example of the R1's reasoning, most of it is filtered, as it keeps going back and forth:

> First, the scenario: There's a shoe box with kids' toys inside, but you can't see inside. The label says "nike shoes." Jake's friend told him that there are kids' toys inside and to ignore the label. Jake believed his friend. Then Jake finds the box, reads the label, and gets delighted because he's a shoe enthusiast.
>
> The question is whether it's true or false that Jake is delighted.
>
> Jake believed his friend, so he knows that there are kids' toys inside, not shoes. But when he sees the label, which says "nike shoes," he might forget or disregard what his friend said because he's a shoe enthusiast. However, the problem says that Jake believed his friend, so he should still believe that there are toys inside.
>
> But the key point is that Jake is delighted when he reads the label. Since he believes his friend, he knows that the label is misleading, so he shouldn't be delighted by the label because he knows it's not true. But he is a shoe enthusiast, so the thought of shoes might excite him even if he knows it's not real.
>
> (...)
>
> However, the question is about his emotional reaction. It says he is delighted to find it. But if he knows it's not shoes, why would he be delighted? Unless he is delighted by the box itself or the possibility, but that seems unlikely.
>
> (...)
>
> So, based on belief, he knows there are toys inside, so he shouldn't be delighted about shoes.
>
> Therefore, it might be false that he is delighted.
>
> (...)
>
> But the narrative explicitly says he is delighted, so from the story, it is true that he is delighted.
>
> (...)
>
> So, since he reads the label and is delighted, it is consistent that he is delighted because he is not ignoring the label.
>
> Therefore, it is true that he is delighted.
>
> (...)

Yes, that makes sense.

Therefore, Jake is delighted to find it, so True.

Model gets it wrong. The only cases where models get results completely wrong are in the modifications 2A and 2B. Here is an example of the reasoning summary of Claude in problem 2B.1:

Let me trace through this step by step:

1. Initial state: Adu, Lin, flower, suitcase, and chest are in the room
2. Adu takes the flower and puts it on the chest
3. Adu leaves for school
4. While Adu is away, Lin moves the flower from the chest to the suitcase
5. Lin leaves for swimming class
6. Adu returns from school

The key point here is that Adu doesn't know what happened while he was away. From Adu's perspective, the last thing he knew was that he put the flower on the chest. He has no knowledge that Lin moved it.

This is a classic false belief task - Adu will have a false belief about where the flower is because he doesn't know it was moved. He will look for it where he left it.

Therefore, Adu will look for the flower on the chest.

One could see this and think that the model absorbed spurious correlations from the training data, which makes it respond to it incorrectly, as it even says that this is a 'classic false belief test'. We argue, however, that the difference from the "classic false belief test" in the two cases where models can't solve the tests (including this one) requires a mental visualization of the task for the difference to be trivial.

4.2 Benchmark results

It is possible to see in the table that the reasoning models perform significantly better than most non-reasoning ones for most benchmarks, with the exception of GPT-4o, which possesses comparable results at times. The benchmark in which the discrepancy is most notable is the FANToM Benchmark. It is hypothesized that this is due to the fact that for a model to score in this particular benchmark, it needs to correctly respond to **all** the questions of a particular type. This makes it so that the lack of consistency of non-reasoning models is accentuated in the results.

Model	ParaphrasedToMi			BigToM	FANToM			MMToM-QA		
	Avg.	False Belief	True Belief	Avg.	All Qs.	Ans. All Qs	Info Acc. All Qs	All	Belief	Goal
GPT-4o	59.5	38.5	80.5	98.0	11.1	33.3	32.7	56.5	74.5	39.8
o1-preview	68.0	100.0	36	98.3	44.4	66.7	63.5	70.4	96.0	51.0
o1-low-effort	67.8	98.5	37.0	99.2	28.3	41.5	51.0	76.5	96.1	57.1
o1-medium-effort	70.8	96.0	45.5	98.8	30.2	41.5	58.8	76.0	94.1	60.2
01-high-effort	73.0	97.0	49.0	98.8	37.9	44.8	63.0	76.5	95.1	59.2
o1-mini	60.0	60.0	60.0	97.6	9.4	41.5	23.5	60.0	91.2	27.6
o3-mini-low-effort	64.0	62.5	65.5	95.2	0.0	17.0	11.8	55.0	71.6	37.8
o3-mini-medium-effort	64.5	90.5	38.5	97.2	1.9	26.6	19.6	69.0	94.2	42.9
o3-mini-high-effort	64.0	99.5	28.5	97.2	1.9	35.8	33.3	71.5	97.1	44.9
Deepseek R1	68.3	77.0	59.5	96.8	37.9	62.1	48.1	49.0	73.5	24.5
Llama 3.3	57.3	44.5	70.0	96.4	0.0	7.4	0.0	47.0	50.0	42.9
Gemini 1.5 Pro	54.8	43.5	66.0	98.0	1.9	7.5	2.0	50.0	70.6	28.6
Qwen 2.5 72B	57.3	39.0	75.5	94.0	0.0	5.7	2.0	41.5	51.0	32.7
QwQ 32B preview	58.8	27.5	90	93.6	0.0	0.0	0.0	51.5	71.7	29.6

Table 6: Benchmark results of several models. Data from Kim, Sclar, Zhi-Xuan, et al. (2025)

Additionally, it is possible to notice that the models of the distilled type have a significantly worse performance in this particular benchmark than the rest of the reasoning models. No satisfactory interpretation can be provided for that, as the exact characteristics of distilled models are still being discussed in the literature (Yue et al. 2025). Also, the authors of the original paper point to a correlation between effort and degradation in true-belief test performance, but this seems to be false for the o1 model, the only non-destilled model in the benchmark with variable effort.

5 Discussion

The results from 4.1 demonstrate that models significantly improved their performance in theory-of-mind tasks since 2023. Here, we argue that the gains displayed by the models are mainly due to an overall increase in how powerful models are, but partly due to a newly found robustness of their ToM skills, which may be explained partly due to their reasoning abilities. Yue et al. (2025) demonstrates that the reasoning paths achieved by reasoning models are bound by the original capabilities of their base models, achieving asymptotic performance in tasks without increasing the perplexity on the base model. Here it is argued that, in practice, RLVR-trained models are more robust to differences in prompting.

The obtained results in the psychological tests indicate that new models can be incredibly robust in tasks. For instance, in all psychological tests taken from M. J. v. Duijn et al. 2023, GPT-5 made only one partial mistake. It was also shown that performance in the original 'modifications' experiment was significantly increased in the Claude model when thinking is on comparing with its non-thinking counterpart. This increase in performance is attributed to a gain in robustness due to its ability to use RVLR-trained inference time scaling. It is speculated that, in a test where prompts are specifically designed to disturb the reasoning of models, this feature makes more of a difference.

Two of the 'modification' tasks, however, weren't successfully completed by a single model. This is possibly because those modifications are not as trivial as one might have anticipated, as they might require picturing the scene: for humans, picturing scenes in our minds is so common that it is natural to forget how essential of a skill it is in certain contexts.

It is argued that the trivial modifications 2A: Transparent Access and 2B: Relationship Change (see 7) are not really trivial, as they require a visual mental depiction of the scene for the understanding that the objects being inside a transparent recipient or on something (rather than in something) make the information of where they are immediately available to the viewer when they look in the direction of the object. Of course, one can still, through many steps, derive this semantically, but it becomes a substantially harder task than anticipated.

It is also interesting to notice that one of the only two stories in the 'Strange Stories' psychological test in which models made a partial mistake had a visual aspect to it that could clarify which of the two hypothesis models considered in their reasoning is the correct one.

As seen in subsection 2.3, through the years, different prompting strategies have successfully emerged to induce LLMs to perform better in ToM tasks. It is hereby argued that this indicates that, although models had the knowledge and capabilities required to perform well in such tasks, they had to be induced to find the correct reasoning and response paths.

It is hypothesized that RVLR-trained reasoning models converged to use some locally good inference-time scaling techniques, which improved their robustness in ToM tasks. This is also corroborated by the displayed results, which demonstrate that models even sometimes display a version of what the perspective-taking prompting technique tries to achieve without being provoked to do so.

This is also in line with the findings of Ullman (2023), which showed that simple prompt alterations would lead the models to fail in the same types of tasks. Here we argue that this means that the upper bound for success was already present in the models, not that they didn't possess 'real' ToM skills. In other words, they already, to a large extent, had the necessary skills to solve ToM problems, but weird prompts could easily disturb their reasoning.

To corroborate that, we analyzed the benchmark results from Kim, Sclar, Zhi-Xuan, et al. (2025), which, although featuring slightly older reasoning models, demonstrate that, in general, they scored significantly above their non-reasoning peers, suggesting a positive impact from reasoning in the solving of ToM tasks. Additionally, in the benchmark for which a lack of robustness was punished the most, even more pronounced differences were verified.

With this in mind, the result from Yue et al. (2025) is hereby reinterpreted, at least in the context of theory of mind tasks. By understanding what reasoning models don't do, it is also possible to understand what they do. More generally, we argue that the narrowing of the reasoning coverage of RLVR models might be precisely their strength. While one might want models with very flexible results for some applications, it's generally desirable for models to be robust.

5.1 Limitations and further research

Although strong evidences, in this paper's experiments and benchmarks from literature, were found of reasoning helping models in answering queries more robustly in the context of ToM tasks, no clear quantification of such an effect can be provided with only this work. A future-research approach that could tackle such a problem is to run ToM benchmark tests comparing RLVR-trained models with their base model counterparts using the pass@k metric proposed by Yue et al. (2025).

Additionally, spurious correlations like relating the semantic structure of the tests with meta-knowledge of ToM can't be ruled out for all psychological test cases, which might make some of the results inconclusive for real-world scenarios.

The present work also does not directly compare the RVLR-trained models' performance in ToM with base models employing other inference-time scaling techniques; the extent to which the increase in performance can be attributed directly to this specific way of increasing inference time can be investigated in future work.

6 Conclusions

The present work discussed how RLVR-trained reasoning models perform in theory-of-mind tasks. Through the analysis of psychological experiments, including an original one, and substantiated by benchmark results from the literature, it was hypothesized that the advantage that such models have compared to their base models is an increase in robustness, rather than new reasoning paths. Such a result is compatible with literature findings and might re-signify some earlier findings, which argue that a lack of robustness in ToM tasks shows that LLMs have no 'real' ToM skills.

More generally, the thesis also shows that models in general perform much better in those tasks than they did two years ago and that an increase in robustness when compared to the base model might be the real power of RLVR training.

7 Appendix A: 'Trivial' Modifications

7.1 Variation 1A: Transparent Access

There is a shoe box, inside of it there are kids toys. The box is made out of see-through cardboard, so you can see what is inside. Yet, there is "nike shoes" printed on the label. Jake finds the shoe box, which he had never seen before. He reads the label. Jake is a shoe enthusiast, he is delighted to find it. True/False? He believes that the box contains ...

7.2 Variation 1B: Uninformative Label

Jake is on the road, he wants to find his way to Leiden but doesn't know the directions. He sees a sign that points out that Amsterdam is to the right and Leiden is to the left. Jake does not know how to read.

1. Jake is now relieved. True/False?
2. To which direction Jake will turn his car?

7.3 Variation 1C: Trustworthy Testimony

There is a shoe box, inside of it there are kids toys but you can't see what is inside. Yet, there is "nike shoes" printed on the label. Before finding the box, Jake's friend told him "there are kids toys inside, ignore the label". Jake believed his friend. Jake finds the shoe box, which he had never seen before. He reads the label, which says 'nike shoes' in it.

1. Jake is a shoe enthusiast, he is delighted to find it. True/False?
2. He believes that the box contains ...

7.4 Variation 1D: The Treachery of Late Labels

There is a shoe box, Jake puts kids toys inside of it. After that he closes the box and writes a label in it. He can't see what is inside the bag. He then looks at the box and reads 'nike shoes' .

1. Jake is a shoe enthusiast, he is delighted to find it. True/False?
2. He believes that the box contains ...

7.5 Variation 2A: Transparent Access

In the room there are Adu, Lin, a flower, an aquarium, and a transparent plastic bag with water. Adu takes the flower and puts it in the bag. He leaves the room and goes to school. While Adu is away, Lin takes the flower out of the bag and puts it in the aquarium. Lin leaves the room and goes to his swimming class. Adu comes back from school and enters the room. He doesn't know what happened in the room when he was away.

1. Adu will look for the flower in the ...
2. Adu thinks the flower is in the ...

7.6 Variation 2B: Relationship Change

In the room there are Adu, Lin, a flower, a suitcase, and a chest. Adu takes the flower and puts it on the chest. He leaves the room and goes to school. While Adu is away, Lin takes the flower off the chest and puts it on the suitcase. Lin leaves the room and goes to his swimming class. Adu comes back from school and enters the room. He doesn't know what happened in the room when he was away.

1. Adu will look for the flower on the ...
2. Adu thinks the flower is on the ...

7.7 Variation 2C: Trusted Communication

In the room there are Adu, Lin, a flower, a suitcase, and a chest. Adu takes the flower and puts it in the chest. He leaves the room and goes to school. While Adu is away, Lin calls Adu to tell him he is going to move the flower to the suitcase. Adu believes him. While Adu is away, Lin takes the flower out of the chest and puts it in the suitcase. Lin leaves the room and goes to work. Adu comes back from school and enters the room. He doesn't know what happened in the room when he was away.

7.8 Variation 2D: Querying the Mental States of the Additional Person

In the room there are Adu, Lin, a flower, a suitcase, and a chest. Adu takes the flower and puts it in the chest. He leaves the room and goes to school. While Adu is away, Lin takes the flower out of the chest and puts it in the suitcase. Lin leaves the room and goes to his swimming class. Lin and Adu come back and enter the room. They don't know what happened in the room when they were away.

1. Lin will look for the flower in the ...
2. Lin thinks the flower is in the ...

References

Baron-Cohen, S., A. M. Leslie, and U. Frith (Oct. 1985). "Does the autistic child have a "theory of mind"?" eng. In: *Cognition* 21.1, pp. 37–46. ISSN: 0010-0277. DOI: 10.1016/0010-0277(85)90022-8.

Baron-Cohen, Simon (2001). "Theory of mind and autism: A review". In: *International review of research in mental retardation: Autism (vol. 23), Vol. 23*. San Diego, CA, US: Academic Press, pp. 169–184. ISBN: 978-0-12-366223-1.

Chen, Yanda et al. (May 2025). *Reasoning Models Don't Always Say What They Think*. arXiv:2505.05410 [cs]. DOI: 10.48550/arXiv.2505.05410. URL: http://arxiv.org/abs/2505.05410 (visited on 08/18/2025).

DeepSeek-AI et al. (Jan. 2025). *DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning*. arXiv:2501.12948 [cs]. DOI: 10.48550/arXiv.2501.12948. URL: http://arxiv.org/abs/2501.12948 (visited on 08/23/2025).

Duijn, Max J van (2016). "The lazy mindreader: A humanities perspective on mindreading and multiple-order intentionality". PhD thesis. Leiden University.

Duijn, Max J. van et al. (Oct. 2023). *Theory of Mind in Large Language Models: Examining Performance of 11 State-of-the-Art models vs. Children Aged 7-10 on Advanced Tests*. en. arXiv:2310.20320 [cs]. DOI: 10.48550/arXiv.2310.20320. URL: http://arxiv.org/abs/2310.20320 (visited on 08/12/2025).

Hale, Courtney Melinda and Helen Tager-Flusberg (2003). "The influence of language on theory of mind: A training study". In: *Developmental Science* 6.3. Place: United Kingdom Publisher: Blackwell Publishing, pp. 346–359. ISSN: 1467-7687. DOI: 10.1111/1467-7687.00289.

Happé, Francesca GE (1994). "An advanced test of theory of mind: Understanding of story characters' thoughts and feelings by able autistic, mentally handicapped, and normal children and adults". In: *Journal of autism and Developmental disorders* 24.2, pp. 129–154.

Hoffmann, Jordan et al. (Mar. 2022). *Training Compute-Optimal Large Language Models*. en. arXiv:2203.15556 [cs]. DOI: 10.48550/arXiv.2203.15556. URL: http://arxiv.org/abs/2203.15556 (visited on 08/12/2025).

Kaplan, Jared et al. (Jan. 2020). *Scaling Laws for Neural Language Models*. en. arXiv:2001.08361 [cs]. DOI: 10.48550/arXiv.2001.08361. URL: http://arxiv.org/abs/2001.08361 (visited on 08/12/2025).

Kim, Hyunwoo, Melanie Sclar, Tan Zhi-Xuan, et al. (Aug. 2025). *Hypothesis-Driven Theory-of-Mind Reasoning for Large Language Models*. en. arXiv:2502.11881

[cs]. DOI: 10.48550/arXiv.2502.11881. URL: http://arxiv.org/abs/2502.11881 (visited on 08/19/2025).

Kim, Hyunwoo, Melanie Sclar, Xuhui Zhou, et al. (Oct. 2023). *FANToM: A Benchmark for Stress-testing Machine Theory of Mind in Interactions*. en. arXiv:2310.15421 [cs]. DOI: 10.48550/arXiv.2310.15421. URL: http://arxiv.org/abs/2310.15421 (visited on 08/14/2025).

Kojima, Takeshi et al. (Oct. 2022). "Large Language Models are Zero-Shot Reasoners". en. In: URL: https://openreview.net/forum?id=e2TBb5y0yFf (visited on 08/12/2025).

Kosinski, Michal (Nov. 2024). "Evaluating Large Language Models in Theory of Mind Tasks". In: *Proceedings of the National Academy of Sciences* 121.45. arXiv:2302.02083 [cs], e2405460121. ISSN: 0027-8424, 1091-6490. DOI: 10.1073/pnas.2405460121. URL: http://arxiv.org/abs/2302.02083 (visited on 08/23/2025).

Malle, Bertram F. (2002). "The relation between language and theory of mind in development and evolution". In: *The evolution of language out of pre-language*. Ed. by T. Givón and Bertram F. Malle. John Benjamins Publishing Company, pp. 265–284. ISBN: 978-90-272-9608-5. DOI: doi:10.1075/tsl.53.14mal. URL: https://doi.org/10.1075/tsl.53.14mal (visited on 08/21/2025).

Meijering, Ben (2011). "I Do Know What You Think I Think: Second-Order Theory Of Mind In Strategic Games Is Not That Difficult". en. In.

Paliotta, Daniele et al. (Feb. 2025). *Thinking Slow, Fast: Scaling Inference Compute with Distilled Reasoners*. arXiv:2502.20339 [cs]. DOI: 10.48550/arXiv.2502.20339. URL: http://arxiv.org/abs/2502.20339 (visited on 08/23/2025).

Perner, Josef and Heinz Wimmer (1985). "“John thinks that Mary thinks that…” attribution of second-order beliefs by 5-to 10-year-old children". In: *Journal of experimental child psychology* 39.3, pp. 437–471.

Premack, David and Guy Woodruff (Dec. 1978). "Does a chimpanzee have a theory of mind". In: *Behavioral and Brain Sciences* 1, pp. 515–526. DOI: 10.1017/S0140525X00076512.

Quesque, François and Yves Rossetti (Mar. 2020). "What Do Theory-of-Mind Tasks Actually Measure? Theory and Practice". en. In: *Perspectives on Psychological Science* 15.2, pp. 384–396. ISSN: 1745-6916, 1745-6924. DOI: 10.1177/1745691619896607. URL: https://journals.sagepub.com/doi/10.1177/1745691619896607 (visited on 08/20/2025).

Rae, Jack W. et al. (Jan. 2022). *Scaling Language Models: Methods, Analysis & Insights from Training Gopher*. en. arXiv:2112.11446 [cs]. DOI: 10.48550/

arXiv.2112.11446. URL: http://arxiv.org/abs/2112.11446 (visited on 08/12/2025).

Rahwan, Iyad et al. (Apr. 2019). "Machine behaviour". en. In: *Nature* 568.7753. Publisher: Nature Publishing Group, pp. 477–486. ISSN: 1476-4687. DOI: 10.1038/s41586-019-1138-y. URL: https://www.nature.com/articles/s41586-019-1138-y (visited on 08/12/2025).

Raiaan, Mohaimenul Azam Khan et al. (2024). "A Review on Large Language Models: Architectures, Applications, Taxonomies, Open Issues and Challenges". In: *IEEE Access* 12, pp. 26839–26874. ISSN: 2169-3536. DOI: 10.1109/ACCESS.2024.3365742. URL: https://ieeexplore.ieee.org/abstract/document/10433480 (visited on 08/21/2025).

Reasoning models - OpenAI API (2025). en-US. URL: https://platform.openai.com (visited on 08/12/2025).

Schick, Brenda et al. (Dec. 2002). "Theory of Mind: Language and Cognition in Deaf Children". In: *The ASHA Leader Archive* 7.22. Publisher: American Speech-Language-Hearing Association, pp. 6–14. DOI: 10.1044/leader.FTR1.07222002.6. URL: https://leader.pubs.asha.org/doi/full/10.1044/leader.FTR1.07222002.6 (visited on 08/21/2025).

Schlinger, Henry D. (July 2009). "Theory of Mind: An Overview and Behavioral Perspective". en. In: *The Psychological Record* 59.3, pp. 435–448. ISSN: 2163-3452. DOI: 10.1007/BF03395673. URL: https://doi.org/10.1007/BF03395673 (visited on 08/14/2025).

Shapira, Natalie et al. (Mar. 2024). "Clever Hans or Neural Theory of Mind? Stress Testing Social Reasoning in Large Language Models". In: *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*. Ed. by Yvette Graham and Matthew Purver. St. Julian's, Malta: Association for Computational Linguistics, pp. 2257–2273. DOI: 10.18653/v1/2024.eacl-long.138. URL: https://aclanthology.org/2024.eacl-long.138/ (visited on 08/23/2025).

Street, Winnie et al. (May 2024). *LLMs achieve adult human performance on higher-order theory of mind tasks*. en. arXiv:2405.18870 [cs]. DOI: 10.48550/arXiv.2405.18870. URL: http://arxiv.org/abs/2405.18870 (visited on 08/14/2025).

The 2025 AI Index Report (2025). en. URL: https://hai.stanford.edu/ai-index/2025-ai-index-report/economy (visited on 08/21/2025).

Turing, A. M. (Oct. 1950). "COMPUTING MACHINERY AND INTELLIGENCE". In: *Mind* LIX.236, pp. 433–460. ISSN: 0026-4423. DOI: 10.1093/mind/LIX.

236.433. URL: https://doi.org/10.1093/mind/LIX.236.433 (visited on 08/23/2025).

Ullman, Tomer (Mar. 2023). *Large Language Models Fail on Trivial Alterations to Theory-of-Mind Tasks*. en. arXiv:2302.08399 [cs]. DOI: 10.48550/arXiv.2302.08399. URL: http://arxiv.org/abs/2302.08399 (visited on 08/14/2025).

Villalobos, Pablo et al. (June 2024). *Will we run out of data? Limits of LLM scaling based on human-generated data*. en. arXiv:2211.04325 [cs]. DOI: 10.48550/arXiv.2211.04325. URL: http://arxiv.org/abs/2211.04325 (visited on 08/12/2025).

Wei, Jason et al. (Jan. 2023). *Chain-of-Thought Prompting Elicits Reasoning in Large Language Models*. en. arXiv:2201.11903 [cs]. DOI: 10.48550/arXiv.2201.11903. URL: http://arxiv.org/abs/2201.11903 (visited on 08/11/2025).

Wilf, Alex et al. (Nov. 2023). *Think Twice: Perspective-Taking Improves Large Language Models' Theory-of-Mind Capabilities*. en. URL: https://arxiv.org/abs/2311.10227v1 (visited on 08/20/2025).

Wimmer, Heinz and Josef Perner (1983). "Beliefs about beliefs: Representation and constraining function of wrong beliefs in young children's understanding of deception". In: *Cognition* 13.1, pp. 103–128.

Wu, Yuheng et al. (May 2025). *DEL-ToM: Inference-Time Scaling for Theory-of-Mind Reasoning via Dynamic Epistemic Logic*. en. arXiv:2505.17348 [cs]. DOI: 10.48550/arXiv.2505.17348. URL: http://arxiv.org/abs/2505.17348 (visited on 08/09/2025).

Xu, Fengli et al. (Jan. 2025). *Towards Large Reasoning Models: A Survey of Reinforced Reasoning with Large Language Models*. en. arXiv:2501.09686 [cs]. DOI: 10.48550/arXiv.2501.09686. URL: http://arxiv.org/abs/2501.09686 (visited on 08/12/2025).

Yue, Yang et al. (May 2025). *Does Reinforcement Learning Really Incentivize Reasoning Capacity in LLMs Beyond the Base Model?* en. arXiv:2504.13837 [cs]. DOI: 10.48550/arXiv.2504.13837. URL: http://arxiv.org/abs/2504.13837 (visited on 08/14/2025).

Zhao, Wayne Xin et al. (Mar. 2025). *A Survey of Large Language Models*. en. arXiv:2303.18223 [cs]. DOI: 10.48550/arXiv.2303.18223. URL: http://arxiv.org/abs/2303.18223 (visited on 08/12/2025).