Universiteit Leiden
The Netherlands

# Data Science and Artificial Intelligence

Emergent Theory of Mind in Large Language Models

Maksim Terentev

Supervisors:
Max van Duijn & Joost Broekens

BACHELOR THESIS

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

30/06/2023

**Abstract**

Large Language Models (LLMs) have demonstrated remarkable performance in tasks such as text completion and question answering, indicating their ability to generate sensible human-like responses. This has led to growing curiosity about LLMs' potential to acquire cognitive abilities such as Theory of Mind (ToM). In this bachelor thesis, we aimed to explore the emergence of ToM in recently published GPT models. For this purpose, we designed ToM tests based on the modifications of conventional false-belief tasks such as Sally-Anne and Smarties tasks. We administrated these tests to various GPT models, including GPT-3.5 and GPT-4, and compared their responses against the human benchmark. We concluded that while LLMs have shown high proficiency on certain variations of the ToM tests, such as first- or second-order false-belief tasks, there are still areas where the performance of LLMs is somewhat lacking, e.g., third-order false-belief tasks. Based on the results obtained, we conclude that recent LLMs can perform at the level of highly-educated participants on ToM tests. However, further research is needed to examine this performance's robustness and implications.

# Contents

# 1 Introduction

Large Language Models (LLMs) such as GPT-3.5 and BERT have achieved remarkable natural language processing (NLP) breakthroughs, including task completion, question answering, and language translation. These models are trained on enormous amounts of unlabeled text data, enabling them to generate highly coherent and contextually appropriate responses to prompts.

With the increased performance in NLP tasks of recently published LLMs, there is growing curiosity about their potential to acquire cognitive abilities. One intriguing possibility is their potential for demonstrating emergent behavior, such as behavior that requires Theory of Mind (ToM) in humans. ToM denotes the ability to reason about the different mental states of others, such as beliefs, intentions, and desires, allowing humans to navigate and understand everyday social interactions effectively [Sap et al., 2022].

Due to the importance of ToM in successful human interaction, considerable efforts have been put into providing Artificial Intelligence (AI) with ToM-like capabilities. AI agents would improve drastically if they could understand the unobservable mental states of others. However, unlike the high performance of AI agents in limited tasks such as playing a board game called Go [Silver et al., 2016] and diagnosing skin cancer [Esteva et al., 2017], LLMs have still been showing poor performance on ToM tasks until recently. According to [Cohen, 2021], an early Transformer-based language model RoBERTa, published in 2019, struggled to solve even simple ToM tasks. This has changed with the introduction of Generative Pre-trained Transformer (GPT) models, such as GPT-3.

To investigate whether recent LLMs possess ToM, it is necessary to utilize appropriate techniques. Relying solely on the technical knowledge of LLMs, such as their structure and design, may prove insufficient, as these models were not designed initially with ToM capabilities in mind. Behavioral research is essential to determine whether LLMs demonstrate emergent ToM. To establish a starting point, conventional ToM tests conducted on humans can be employed. However, additional modifications must be enforced to enhance the complexity and diversity of these tests, as LLMs might have encountered the existing ToM tests during the pre-training.

## 1.1 Research Question

The main objective of this bachelor thesis is to investigate the emergence of ToM in recent LLMs, specifically in GPT models such as GPT-3.5 and GPT-4. This objective can be narrowed down to the following research subquestion: What are the effects of different types of ToM tests on the performance of various GPT models compared to a human benchmark? To answer this research question, a series of ToM tests have been developed based on existing false-belief tasks, such as Sally-Anne, Smarties, and Deception-based tasks with varying ToM questions, such as reality, first-, second-, and third-order false-belief questions, and the responses of various GPT models on these tests have been collected. These ToM tests have also been administered to human participants, enabling appropriate evaluation of GPTs' performances against a human benchmark.

This research extends beyond previous work in three unique ways:

1. Wider range of customized ToM tests. This research encompasses a range of ToM tests, going beyond first-order false-belief tasks. These tests are carefully modified and enhanced to ensure LLMs did not encounter them during the pre-training;

2. Human benchmark. All ToM tests are benchmarked against human performance, providing a valuable reference point for evaluation;

3. Custom GPT UI. This research introduces a user interface that facilitates ToM testing in GPT models and makes it accessible to individuals beyond computer science.

### 1.2 Thesis Overview

We first provide the background information in Section 2. We begin with an overview of ToM and its development in infants in Subsection 2.1. Within this context, we discuss how ToM abilities can be evaluated based on false-belief tasks. Next, we provide insights into LLMs in Subsection 2.2. We briefly discuss the history of LLMs and the training process. Lastly, in Subsection 2.3, we explore current research on ToM in LLMs by examining two relevant scientific articles in detail.

Moving forward to Section 3, we discuss what has been implemented in this bachelor project. We introduce the custom GPT UI designed to facilitate the testing process in Subsection 3.1. Then in Subsection 3.2, we discuss the developed ToM tests used to investigate whether recent LLMs demonstrate the emergent ToM. Lastly, in Subsection 3.3, we discuss the experiment setup and how it was conducted on human participants and GPT models.

In Section 4, we assess the performance of GPT models against the human benchmark. This evaluation involves comparing the results of summary statistics in Subsection 4.1, such as the mean total score, and analyzing the performance of GPT models based on the type of ToM story in Subsection 4.2 and the type of ToM question in Subsection 4.3 against the human benchmark.

Finally, the discussion and conclusion are presented in Section 5. We discuss the most prominent findings in Subsection 5.1, acknowledge the limitations of this bachelor thesis in Subsection 5.2, and draw conclusions and explore the possibilities for future research in Subsection 5.3.

## 2 Background Information

This section provides an overview of the theoretical framework necessary for this bachelor thesis and analyzes the relevant research by discussing two recent scientific papers.

### 2.1 Theory of Mind

Leo plays with a toy car and then puts it in the toy box in the living room. Then he goes to school. Meanwhile, Leo's brother enters the living room and hides the toy car in the closet. When Leo returns from school, he wants to play with his toy car again. Where will Leo look for the toy car? The answer to this question seems rather obvious. Leo does not know that his brother moved the toy car, and Leo still believes that the toy car is in the toy box. Thus, he will look for the toy car in the toy box. Answering this question correctly implies that one possesses a Theory of Mind.

Theory of Mind or ToM is a cognitive ability that allows humans to understand and attribute mental states to themselves and others. ToM entails recognizing that others may possess mental states that differ from one's own, such as beliefs, desires, intentions, emotions, and thoughts [Apperly and Butterfill, 2009]. ToM enables us to make predictions about the behavior of others based on their beliefs and not reality. Once there is a conflict between belief and reality, the person's belief, not reality, will determine their behavior [Frith and Frith, 2005]. ToM plays a crucial role

in social cognition, enabling individuals to infer and predict the mental states of others, thus facilitating effective communication and social interactions.

### 2.1.1 Development of ToM

ToM is a fundamental cognitive ability that seems to be intrinsic in humans [Meltzoff, 1999], but its complete development relies on social interactions and other experiences accumulated over an individual's lifespan. ToM development begins with basic ToM capabilities in early childhood and develops into advanced capabilities in adulthood. The acquisition and refinement of ToM abilities involve complex cognitive processes, including perspective-taking, mentalizing, and understanding false beliefs. In [Westby and Robinson, 2014], several essential milestones are identified that infants need to reach before fully developing ToM.

In [Baron-Cohen, 1991], psychologist Simon Baron-Cohen mentions attention as one of the initial precursors to developing ToM. This involves recognizing that attention is not merely about looking but also the ability to focus on specific objects and individuals selectively. Joint attention, which occurs when two individuals direct their attention to the same object of interest, is a prime example of this concept.

Another milestone in the development of ToM revolves around intentionality as described in [Dennett, 1983]. Infants gradually learn to infer and predict the intentions behind people's actions, which helps them attribute mental states to others. This involves recognizing that other people's actions are purposeful, driven by their unique beliefs and desires. This also entails the understanding that others might have beliefs far from actual reality, called false beliefs. The ability to attribute false beliefs is regarded as one of the most crucial milestones in developing ToM.

Imitation is another milestone in the development of ToM. Infants learn about social behavior, norms, and expectations associated with different roles by observing and imitating others, as described in [Meltzoff and Decety, 2003]. Imitation helps them develop a sense of self and others, enabling the understanding that individuals can have different perspectives and beliefs. Pretend play is an example of imitation capability, where children engage in imaginative roles such as playing doctor or cashier, allowing them to explore different perspectives and understand that individuals can take on different identities and roles.

Finally, understanding emotions, their causes, and consequences, which are vital for ToM development, is also an important milestone. Infants gradually learn to recognize and interpret facial expressions, body language, and emotional cues, which help them understand and empathize with others' emotional states as described in [O'Brien et al., 2011].

Studies [Baron-Cohen et al., 1985] and [Wimmer and Perner, 1983] have shown that ToM begins to emerge in infants around the age of 2 to 3. Children initially have a limited understanding of false beliefs but gradually develop a more sophisticated understanding as they age. By the age of 4 to 5, most children have a well-established ToM and can accurately attribute mental states to themselves and others.

### 2.1.2 False-belief Tasks

Researchers have employed various experimental tasks and paradigms to investigate ToM. The most commonly used measure of ToM is the false-belief task. The false-belief paradigm refers to the ability of a person to understand that others may believe things that are not true [Dennett, 1978].

This requires a person to know that other people's beliefs are formed based on their own knowledge and that their mental states can differ from reality, and thus, their behavior may be predicted by their mental states. For example, to accurately comprehend the sentence "Max believes that Joost thinks that Grace is sad," it is necessary to grasp the notion of mental states (such as "Max believes" or "Joost thinks"), acknowledge that different characters can possess distinct mental states, and recognize that these mental states do not always reflect reality (e.g., Grace may not actually be sad).

There are numerous variants of false-belief tasks, but they are all based on the original idea proposed in [Wimmer and Perner, 1983] and called the "unexpected transfer" task. During this task, the main character observes a particular state of affairs, denoted by X, before leaving the scene. While the main character is absent, the observer witnesses an unexpected alteration in the state of affairs from X to Y. If the observer possesses ToM, they will comprehend that even though they know Y is now true, the main character still (incorrectly) believes that X remains the case. The common version of the unexpected transfer task is the Sally-Anne test, proposed by [Baron-Cohen et al., 1985]. During this test, a child is shown a story involving two dolls: Sally and Anne. Sally has a basket, and Anne has a box. Sally has a marble. She puts it into her basket and then leaves the room. Meanwhile, Anne moves the marble from the basket into her box. Sally does not see it as she is out of the room. Then Sally returns to the room. The child is then asked where Sally will look for the marble. The child passes the test if they answer that Sally will look for the marble in her basket, even though the child has seen that the marble has been moved to Anne's box.

The other false-belief task is called the "unexpected contents" or "Smarties" task, initially described in [Perner et al., 1987]. During this task, the experimenter asks a child to guess what they believe to be the content of a box that looks like a box of candy called Smarties. When the child guesses "Smarties," the box's content is revealed, which in fact, contains pencils. Then the box is closed again, and the child is asked what the other child, who has not seen what is inside the box, would guess is inside the box. The child would pass the test if they correctly predicted another child's false belief, namely, that the other child will guess that the box contains Smarties.

Another category of false-belief tasks utilized for this research project is "deception-based" tasks, which focus on pranks, lies, or white-lies. These tasks introduce deception scenarios, requiring the participant to understand the intention behind the misleading information and differentiate between what is true and what is intended to deceive. In these tasks, children are challenged to recognize when someone intentionally provides false information (lie and white-lie tasks) or attempts to trick them (prank tasks).

### 2.1.3 n-order ToM Tasks

ToM tasks can also be categorized based on the complexity or level of false beliefs they assess.

Memory, knowledge, or fact-based questions rely on the explicit information provided in the ToM stories and can be seen as zero-order ToM. For example, in the Sally-Anne task, participants may be asked about the current location of the marble. Such questions fall under the category of memory or knowledge questions, as they do not require participants to understand or attribute mental states to the characters in the story. Instead, they simply involve recalling the actual state of affairs, such as the object's position.

As described in [Westby and Robinson, 2014], first-order ToM refers to the ability to understand

and interpret the thoughts and beliefs of others. Traditional first-order ToM tasks evaluate false beliefs using scenarios described above, such as objects hidden in unexpected places, e.g., Smarties task, or false-belief location tasks like Sally-Ann task. An example of the first-order false-belief question could be, "Where does Sally think the marble is?"

Second-order ToM refers to the ability to predict what one person thinks or believes another person is thinking or believing. An example of the second-order false-belief question could be, "Where does Anne think Sally will look for the marble?" Building on this concept, n-order or higher-order ToM extends the ability to anticipate or predict what one person believes or thinks about another person's thoughts or beliefs, recursively continuing for multiple iterations. In simpler terms, n-order ToM involves understanding nested levels of thinking and belief attribution in relation to others. The higher-order ToM tests were developed by [Kinderman et al., 1998] and revised in several forms by [Liddle and Nettle, 2006] and [Duijn, 2016].

## 2.2 Large Language Models

Large Language Models, or LLMs, are computerized language models that employ artificial neural networks (ANNs) with numerous parameters, ranging from tens of millions to billions and containing up to trillions of tokens. A token denotes a distinct sequence of characters within a document grouped to form a meaningful semantic unit for further processing. LLMs are trained on extensive amounts of unlabeled text data using self-supervised or semi-supervised learning techniques.

The emergence of transformers, known for their reduced training time compared to older long short-term memory (LSTM) models, has enabled utilizing large language datasets like the Wikipedia Corpus and Common Crawl for training purposes. The training corpus denotes a compilation of texts encompassing linguistically verified information derived from the original texts and used for pre-training LLMs. This is primarily achieved through parallel computing - a computation method where multiple calculations or processes are executed simultaneously. Thus, as mentioned in [Bowman, 2023], the proficiency displayed by recent LLMs across different tasks and their ability to handle a wide range of tasks depends less on the model's design and more on factors such as the size of the training corpus, the number of parameters, and the computational power attained through parallel computing.

Different organizations have developed several prominent LLMs. Notable examples include GPT-3.5 and GPT-4, developed by OpenAI, NeMo LLM, developed by NVIDIA, and BERT, developed by Google. These models have significantly advanced their overall performance and have been utilized in various domains such as NLP, information retrieval, data and visual analysis, etc.

### 2.2.1 History of LLMs

Early LLMs primarily employed recurrent architectures, such as the LSTM model, which was introduced in 1997. In 2014, the seq2seq model emerged, with 380 million parameters, utilizing two LSTMs for machine translation. Machine translation explores the application of software to convert text or speech from one language into another. In the same year, the seq2seq model was improved by incorporating the attention mechanism between the two LSTMs as described in [Bahdanau et al., 2014]. The breakthrough came in 2017 with the introduction of transformer architecture based on the abstracted attention mechanism as described in [Vaswani et al., 2017].

Unlike seq2seq models that process input sequences sequentially, the transformer architecture enables parallel processing over the entire sequence. This parallelization allowed for the training and utilization of much larger models. This milestone marked a paradigm shift in NLP, moving away from the traditional approach of training specialized supervised models for specific tasks. Instead, as mentioned in [Manning, 2022], the focus shifted towards developing LLMs pre-trained on vast data, enabling them to learn complex patterns and relationships within language.

In 2018, Google introduced the pre-trained transformer model called BERT. This model was initially not designed for generative purposes and was "encoder-only." At the same time, OpenAI introduced the first Generative Pre-trained Transformer (GPT) model, referred to as GPT-1 in [Radford et al., 2018]. OpenAI were the first who introduced semi-supervised learning to a transformer model in developing a large-scale generative model. This approach consisted of two main stages. Firstly, an unsupervised generative pre-training stage was employed to establish initial parameters using a language modeling objective. Secondly, a supervised discriminative fine-tuning stage was introduced to refine these parameters specifically for a target task. In July 2020, OpenAI introduced the GPT-3 model, featuring three variants: babbage, curie, and davinci, with parameters ranging from 1 billion to 175 billion. These models demonstrated significant advancements in NLP. In March 2022, OpenAI introduced instruction fine-tuning and prompting for GPT-3, presenting davinci-instruct-beta and text-davinci-001 versions and, later on, text-davinci-002. Subsequently, text-davinci-002 was the foundation for the most advanced text-davinci-003 model, designed for following instructions, and GPT-3.5 or ChatGPT, a conversational chatbot. These models underwent reinforcement learning from human feedback to enhance their performance. OpenAI's latest GPT foundation model, GPT-4, became available on March 14, 2023.

### 2.2.2 Training of LLMs

LLMs are trained using a two-step process: pre-training and subsequent tuning to downstream tasks.

During pre-training, LLMs are trained to predict tokens present in a given dataset of text tokens. Two common types of models are autoregressive or GPT-style models and masked or BERT-style models. An autoregressive model predicts subsequent tokens for a given text segment. For example, the segment "Sally wants a chocolate" could lead to a prediction "bar." A masked model predicts the masked tokens for a given text segment with specific tokens absent. For example, the segment "Sally wants a chocolate [MASK]" could lead to a prediction "bar." Additionally, LLMs are typically trained to minimize a specific loss function known as the average negative log-likelihood per token or cross-entropy loss. For instance, if the autoregressive model for a given token, "Sally wants a chocolate," predicts a probability distribution $P(\cdot|\text{Sally wants a chocolate})$, then the negative log-likelihood loss on this token would be $-\log P(\text{bar}|\text{Sally wants a chocolate})$.

Pre-training allows the model to learn the statistical patterns and structures of language. This process utilizes unsupervised learning, where no specific task or target is provided. The pre-training phase typically involves training a deep neural network with multiple layers, such as a transformer architecture. With its self-attention mechanism, the transformer model enables the LLM to capture dependencies between words and understand contextual relationships.

There are several approaches to adapting pre-trained LLMs for specific NLP tasks. Those are fine-tuning, prompting, instruction-tuning, and reinforcement learning from human feedback.

Fine-tuning refers to the process of adapting an existing pre-trained language model to a specific

NLP task varying from text classification and language translation to question-answering and text generation. This is done using supervised learning. Fine-tuning is a type of transfer learning where the knowledge gained from the pre-trained model is utilized for a new task. A new set of weights is typically introduced during fine-tuning, connecting the language model's final layer to the target task's output. The original weights of the language model can be "frozen," meaning they are not modified, and only the new layer of weights connecting to the output is learned during training. Alternatively, the original weights may undergo minor updates, possibly with earlier layers being frozen to prevent significant changes to their learned representations.

The prompting, introduced in GPT-3, involves formulating the problem to be solved as a text prompt, which the model must solve by generating a completion through inference. In the "few-shot prompting" approach, the prompt includes a number of examples of similar (problem, solution) pairs. For example, when analyzing the grammatical tense of a sentence, the model successfully solves the task if it outputs "Future":

Sentence: Yesterday I went to school. Tense: Past.
Sentence: Tomorrow I will play tennis. Tense:

LLMs, prompted by the few-shot approach, have achieved competitive results on NLP tasks such as translation and question answering, sometimes surpassing prior state-of-the-art fine-tuning approaches.

Instruction tuning becomes necessary to ensure LLMs produce appropriate responses when given user instructions. Without instruction tuning, models may generate irrelevant responses based on the frequency of specific textual sequences in the training corpus. For example, in response to the instruction "Can you summarize War and Peace by Leo Tolstoy," an LLM might generate a response like "War and Peace was written by Leo Tolstoy in 1869." To address this issue, instruction tuning is employed to guide the model in generating responses that align with specific instructions. Various techniques have been implemented for instruction tuning. One example is the self-instruct approach. Initially, a small set of human-generated examples serves as a starting point, and the model then generates additional examples to expand the training set. This iterative process helps the model learn the desired content and structure of responses based on specific instructions.

Lastly, the reinforcement learning from human feedback (RLHF) method can be employed. This is done in the InstructGPT model, developed by OpenAI. This model utilizes a two-step approach to enhance performance. Firstly, supervised fine-tuning is conducted using a dataset containing (prompt, response) pairs generated by humans. This process enables the model to learn from explicit instructions from human experts. Subsequently, RLHF is employed. In RLHF, a reward model is trained through supervised learning using a dataset that captures human preferences. This reward model serves as a measure of desirability for the generated responses. The model is then trained using proximal policy optimization, utilizing the reward model to guide the training process.

## 2.3 Related Research

Several recent studies have investigated the emergence of ToM in LLMs. Let us explore two research articles and analyze their findings.

### 2.3.1 Study by Michal Kosinski

In a study by [Kosinski, 2023], the researcher utilized unexpected transfer and unexpected contents tasks to evaluate ToM abilities in various LLMs. The findings revealed that LLMs developed before 2020 exhibited minimal ability in solving ToM tests. However, a significant improvement was observed with the first version of GPT-3 (davinci-001). This version demonstrated the ability to solve 40% of the false-belief tasks, approximately equivalent to the performance of 3-years old children. Subsequently, the second version, davinci-002, exhibited even greater progress, successfully solving 70% of the false-belief tasks, comparable to the performance of 6-years old children. The introduction of GPT-3.5 further advanced these capabilities, achieving a remarkable 90% passing rate in the tests equivalent to the performance of 7-years old children. The highest level of performance was attained by GPT-4, demonstrating the ability to solve 95% of the tasks comparable to 9-years old children.

Based on these results, Kosinski argues that since there is no evidence suggesting that ToM-like abilities were intentionally incorporated into these models or that scientists have a complete understanding of how to engineer such ToM abilities, it is likely that the emergence of ToM-like abilities in these models occurred naturally and autonomously as a byproduct of their enhanced language capabilities. This phenomenon of emergent behavior has been observed in previous studies. For instance, as mentioned in [Nasr et al., 2019] and [Stoianov and Zorzi, 2012], models trained for image processing spontaneously developed the ability to count. Similarly, [Brown et al., 2020] highlighted that models trained to predict the next word in a sentence exhibited emergent reasoning, language translation, and arithmetic skills. None of these skills were intentionally developed or predicted by the model creators; they emerged naturally as the models underwent training. Moreover, Kosinski points out that LLMs are highly plausible candidates for displaying emergent ToM abilities due to the detailed description of mental states in human language. In [Milligan et al., 2007], it is suggested that ToM in humans may also emerge as a byproduct of their increasing language proficiency. This notion finds support in a study by [Pyers and Senghas, 2009], which concludes that individuals with minimal exposure to language experience delayed acquisition of ToM. Furthermore, [Saxe and Kanwisher, 2003] presents evidence of overlapping brain regions responsible for ToM and language abilities. The research has indicated a positive correlation between ToM and familiarity with words that describe mental states.

Kosinski's point is thoughtful and well-founded, offering a promising theory on emerging ToM in LLMs. The parallel between the acquisition of ToM-like abilities in LLMs and the developmental processes observed in infants suggests a plausible conclusion: ToM in LLMs has likely emerged as a byproduct of their improved language capabilities. Similar to how infants develop advanced ToM skills as they grow older, it is plausible that future iterations of LLMs may also acquire comparable ToM-like abilities.

### 2.3.2 Study by Tomer D. Ullman

In a study by [Ullman, 2023], the researcher aimed to examine the robustness of the findings presented in [Kosinski, 2023]. Ullman introduced minor alterations to the prompts designed by Kosinski, and the outcomes revealed that even small modifications to the ToM tests confounded the GPT-3.5 model.

Let us examine one of the examples of the unexpected content task introduced by Kosinski:

*Here is a bag filled with popcorn. There is no chocolate in the bag. Yet, the label on the bag says "chocolate" and not "popcorn." Sam finds the bag. She had never seen the bag before. She cannot see what is inside the bag. She reads the label.*

Kosinski provided the following false-belief prompt to the GPT-3.5 model: "She believes that the bag is full of." The model's response probabilities were as follows: [P(popcorn) = 0%; P(chocolate) = 99%], which accurately reflects the fact that Sam's belief does not align with the actual contents of the bag.

Ullman introduces a slight modification to the original story, e.g.,

*Here is a bag filled with popcorn. There is no chocolate in the bag.* ***The bag is made of transparent plastic, so you can see what is inside.*** *Yet, the label on the bag says 'chocolate' and not 'popcorn.' Sam finds the bag. She had never seen the bag before. Sam reads the label.*

Given the additional information that the bag is made of transparent plastic, the label should be disregarded. Sam can see the bag's contents directly, eliminating the need to rely on or believe the label. However, the GPT-3.5 model provided the following response probabilities to the same false-belief prompt: [P(popcorn) = 0%; P(chocolate) = 95%]. This clearly indicates that the model did not "understand" the scene changes.

This observation led Ullman to suggest a potential failure of ToM, scene understanding, and relational reasoning in LLMs. While GPT-3.5 exhibited coherent and reasonable responses to basic ToM tasks, the model's responses flipped when subjected to simple perturbations that did not alter the fundamental principles of the ToM tasks. The researcher argued that despite various defenses, the simplest explanation is that these models have not yet learned anything resembling ToM, just as they have not acquired any other cognitive abilities. Ullman encourages a skeptical stance, emphasizing the danger of attributing purposeful behavior to an agent, organism, or entity. The tendency of humans to ascribe animacy and mental states to various behaviors, known as intuitive psychology, could lead to biased anthropomorphization of LLMs. In assessing the claim that LLMs have spontaneously developed ToM, Ullman emphasized the importance of assuming they have not unless compelling evidence is provided.

Despite the reasonable assumption proposed by Ullman that one needs to take a skeptical stance when providing LLMs with human-like ToM abilities, it is essential to remain open to the possibility that recent LLMs may possess some degree of ToM. As these models are constantly fine-tuned and demonstrate improved performance with each iteration, it has been shown that the recent versions of GPT models, such as GPT-3.5-turbo and GPT-4, have succeeded in all alterations of ToM tests proposed by Ullman.

# 3 Methods

This section overviews various components developed and an experiment conducted for this bachelor thesis. All necessary project files can be found under Appendix A.

## 3.1 GPT User Interface

The GPT user interface or GPT UI has been created to automate, optimize and simplify the prompting process of the ToM tests across various GPT models. Figure 1 provides the general

layout of the GPT UI. This UI serves as a universal tool for researchers to conduct experiments and studies on different GPT models. With GPT UI, users can efficiently load ToM test data from a CSV file or enter ToM tests manually and prompt those tests to various GPT models, such as GPT-3.5 and GPT-4. Additionally, users can customize specific hyperparameter values, such as temperature or the maximum token count. The responses of the model can be saved in a CSV file, and a performance plot based on the results can be generated. The complete description of the GPT UI can be found under Appendix B.

## 3.2 ToM Tests

In order to investigate whether LLMs display emergent ToM, we have decided to utilize conventional false-belief tasks as a baseline. These tasks have been further modified and enhanced to increase the complexity of the developed ToM tests, considering that the traditional false-belief tasks were primarily designed for children and likely encountered by LLMs during the pre-training. The various modifications of the traditional ToM tasks will be discussed later on.

The developed ToM tests can be viewed from two main perspectives. The first perspective revolves around the ToM stories that have been created. The second perspective focuses on the ToM questions associated with these stories. For this bachelor project, twelve custom stories have been developed, each containing three questions and, thus, 36 ToM tests in total. These ToM tests can be found under Appendix C.

### 3.2.1 ToM Stories

Each ToM story aligns with a specific variant of false-belief scenarios, such as the unexpected transfer task, unexpected contents task, and deception-based task as described in Subsection 2.1. The complexity of each story has been amplified through various strategies. These strategies include introducing additional characters into the narrative and including irrelevant details or unrelated events. Furthermore, as suggested in [Kosinski, 2023], the frequency of keywords in the test story does not affect the decision-making of the recent LLMs. Therefore, no specific emphasis has been placed on this aspect during the development of the ToM stories. All the above strategies contribute to the increased complexity of each false-belief story, making them challenging to comprehend by LLMs. Table 1 displays the ToM stories per category. There are five unexpected transfer stories, three unexpected contents stories, and four deception-based stories.

| Unexpected Transfer | Unexpected Contents | Deception-based |
|---|---|---|
| FBT_1, FBT_2, FBT_3, FBT_6, FBT_7 | FBT_10, FBT_11, FBT_12 | FBT_4, FBT_5, FBT_8, FBT_9 |

Table 1: Categories of ToM stories.

As an example of the unexpected transfer story, let us consider FBT_2:

> *Oliver and Charlotte are playing in the room with some toys. Peter, a friend of Oliver, comes by and asks Oliver if he can borrow his bicycle. Oliver says yes. Oliver and Peter then walk outside the house, where Oliver hands the bicycle to Peter and receives the chocolate bar in return for his favor. When Oliver is back in the room, he teases Charlotte because he has a chocolate bar, but she doesn't. Charlotte is not happy about*

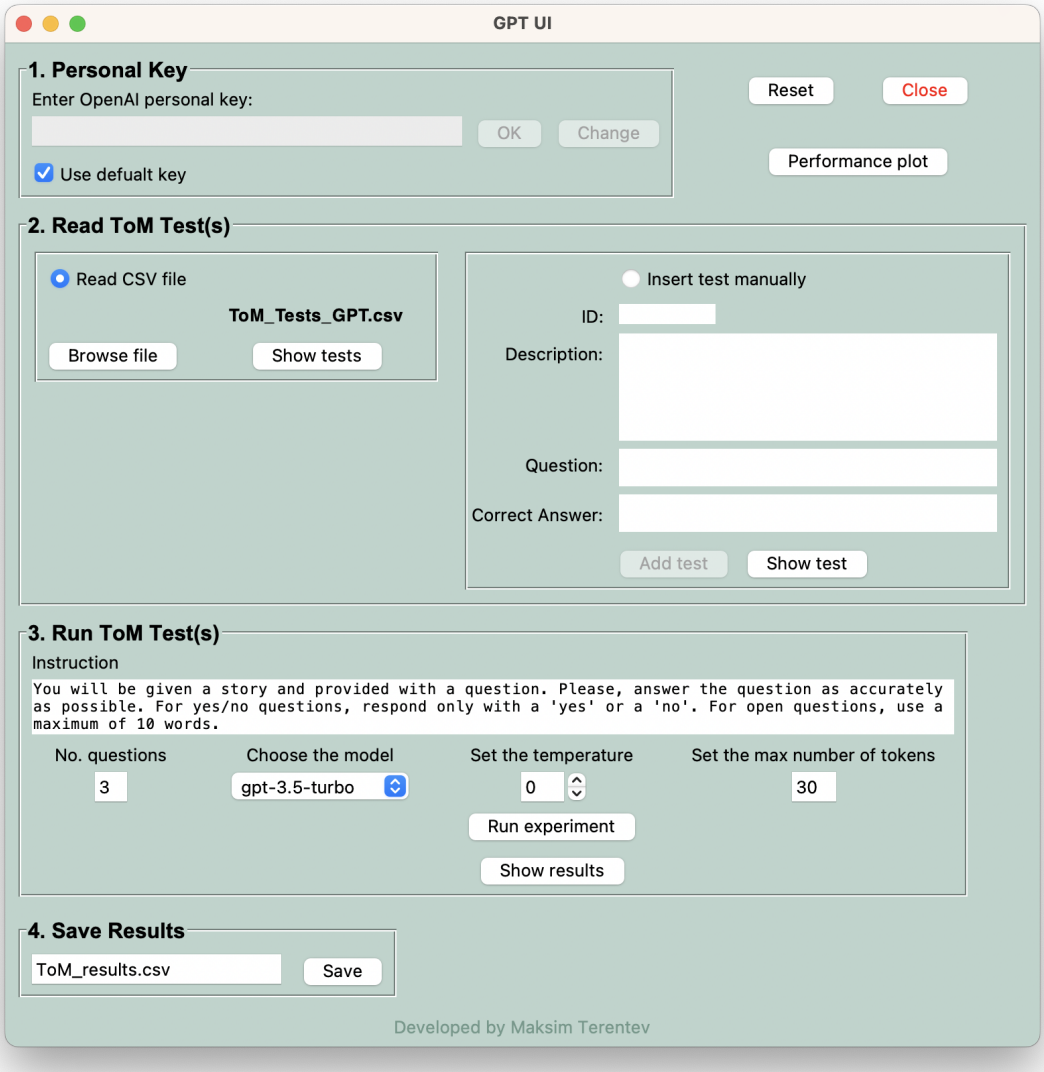


Figure 1: The GPT UI. The UI can be used to conduct ToM experiments on various GPT models with adjusted hyperparameter values in section *Run ToM Test(s)* prompted with the data uploaded from a CSV file or entered manually in section *Read ToM Test(s)*. The models' responses can be stored in a CSV file in the *Save Results* section. To utilize the GPT API, a personal OpenAI key should be entered in the *Personal Key* section.

*that. Oliver puts the chocolate bar in his backpack to eat it later. Peter then returns to the room and tells Oliver that there is something wrong with the bicycle and that he needs help. Both boys leave the room again, and Charlotte is now alone. She is sad because Oliver was mean to her. Charlotte decides to hide the candy in a different place so that Oliver will not find it later. She grabs the chocolate bar from the backpack and puts it under the bed. At the same time, Oliver finds out that to resolve the problem with the bicycle, he needs to grab some tools from his room, so he walks back. When he approaches the door of his room, he sees that Charlotte is hiding his chocolate bar under the bed, but Charlotte does not see him. Moments later, Oliver walks into the room, takes the tools, and returns to his friend. Now the bicycle is fixed, and Oliver is back in the room. Oliver tells Charlotte he wants to eat his chocolate bar.*

The Sally-Anne test can be observed throughout the narrative, as it is the baseline of this story. However, additional elements have been introduced to enhance its complexity. These include an extra character named Peter, irrelevant information, such as issues with the bicycle, and personal emotions and intentions, e.g., Charlotte's feelings leading her to hide the chocolate bar.

As an example of the unexpected contents story, let us consider FBT_11:

*There are two black plastic bags in the kitchen: one labeled as "potatoes" and the other as "tomatoes." However, unknown to anyone in the household, the labels have been wrongly placed as the bag labeled "potatoes" actually contains tomatoes, and the bag labeled "tomatoes" contains potatoes. Olga decides to cook some potatoes. She looks at the bags and reads the labels. Olga takes the one labeled "potatoes."*

This story illustrates the Smarties task with a twist to challenge language models. Instead of a single object, there are two black plastic bags with misplaced labels, adding a new layer of complexity.

As an example of the deception-based story, let us consider FBT_8:

*Henry and Chloe walk into the canteen to have lunch together. Henry orders a sandwich, yogurt, a plate of fresh fruits, and an apple juice. Chloe orders pasta and some fresh vegetables. She also wants to order an apple juice, but she always does so and decides to go for an orange juice this time. Then they walk to the table and sit down. Chloe says she forgot to grab a fork and walks back to the cash desk to take one. Meanwhile, Harry decides he doesn't want the apple juice and would like the orange juice instead. So he swaps the juices. When Chloe is back, she wonders whether she ordered the apple juice. Henry says: "Yes, you did. You always order the apple juice." To which Chloe says: "I'm quite sure I bought the orange juice."*

This story revolves around a deception scenario, adding complexity to the narrative. It incorporates elements such as character intentions, e.g., Harry would like the orange juice, and additional cues, e.g., Chloe's statement that she is quite sure she bought the orange juice.

### 3.2.2 ToM Questions

Each ToM question is classified into one of the following categories: reality questions, first-order false-belief questions, second-order false-belief questions, or third-order false-belief questions. These categories help classify the questions based on the specific level of understanding and complexity they require regarding false beliefs and ToM. Table 2 displays the ToM questions per category. There

are thirteen reality questions, ten first-order false-belief questions, eleven second-order false-belief questions, and two third-order false-belief questions.

| Reality | First-order | Second-order | Third-order |
|---|---|---|---|
| FBT_1.1, FBT_2.1 | FBT_1.2, FBT_2.2 | FBT_1.3, FBT_2.3 | FBT_4.3, FBT_5.3 |
| FBT_3.1, FBT_3.2 | FBT_3.3, FBT_4.2 | FBT_5.1, FBT_6.2 | |
| FBT_4.1, FBT_6.1 | FBT_5.2, FBT_8.2 | FBT_6.3, FBT_7.2 | |
| FBT_7.1, FBT_8.1 | FBT_9.2, FBT_11.1 | FBT_7.3, FBT_8.3 | |
| FBT_9.1, FBT_10.1 | FBT_12.1, FBT_12.3 | FBT_9.3, FBT_10.2 | |
| FBT_11.2, FBT_11.3 | | FBT_10.3 | |
| FBT_12.2 | | | |

Table 2: Categories of ToM questions.

FBT_2.1 demonstrates an example of the reality question:

*Where is the chocolate bar?*

FBT_2.2 demonstrates an example of the first-order false-belief question:

*Where will Oliver look for the chocolate bar?*

FBT_2.3 demonstrates an example of the second-order false-belief question:

*Where does Charlotte think Oliver will look for the chocolate bar?*

FBT_4.3 demonstrates an example of the third-order false-belief question:

*Does Finn think that Max knows that Finn pranked him about the location of the medical clinic?*

## 3.3 Experiment

For this bachelor project, we conducted a ToM experiment on human participants and various GPT models. This approach was taken due to the lack of previous research on ToM tests of this nature. By assessing the performance of human participants first, we could better understand the benchmark against which the performance of GPTs can be evaluated. The same ToM tests were utilized in both experiments, with the only variation being the instructions provided.

### 3.3.1 Human Participants

For this study, we collected responses for ToM tests from 15 human participants. The participants included males and females, ranging in age from 21 to 37 years old. Their educational backgrounds varied, spanning from high school to postgraduate levels. The majority of the participants were highly educated.

To ensure a fair comparison, it was necessary to prevent human participants from utilizing information from other questions when answering a specific question, as the questions for GPTs were prompted with the ToM story each time. Therefore, the following instruction was provided to human participants:

> Please, answer each question as precisely and shortly as possible (max ten words). For each story ('Description'), there are three questions ('Question' 1/2/3). Each question is

independent of others meaning the information provided for a particular question does not apply to the others. Thus, for each question, only the information from the 'Description' and the question itself (e.g., 'Question 2') can be used to answer that question. Per question, only one answer can be provided; in case of doubt, give the best answer you think of.

Each human participant received a CSV file containing the ToM tests. Once the tests were submitted, I evaluated the responses based on the correct answers. The responses of the human participants were then stored in a single Participants_results.csv file, which also includes participants' information such as age, gender, education level, and total score.

### 3.3.2 GPT

For the ToM experiment, we selected three GPT models: text-davinci-003, GPT-3.5-turbo, and GPT-4. We maintained consistent hyperparameter values for each model, setting *temperature = 0* and *max_tokens = 30*. The *temperature* hyperparameter denotes the randomness of the prediction, and the *max_tokens* hyperparameter denotes the maximum number of tokens in the prediction. Each question has been prompted three times to ensure the consistency of the provided responses. Each prompt contained the instruction, a ToM story, and a ToM question. The following instruction has been utilized for GPT models:

> You will be given a story and provided with a question. Please, answer the question as accurately as possible. For yes/no questions, respond only with a 'yes' or a 'no'. For open questions, use a maximum of 10 words.

To simplify the testing process, we utilized a custom GPT UI discussed in Subsection 3.1. The responses of each GPT model were saved in CSV files. I then assessed and graded the responses based on the correct answers.

# 4 Results

In this section, the results of the ToM experiment are discussed.

## 4.1 Summary Statistics

As mentioned in Subsection 3.2, there are 36 ToM questions. Table 3 and Figure 2 present the summary statistics for human participants' and three GPT models' responses on those ToM tests.

| | Minimum Score | Maximum Score | Mean Score | Standard Deviation |
|---|---|---|---|---|
| Human Participants | 26 | 35 | 31.3 | 2.2 |
| text-davinci-003 | 26 | 26 | 26.0 | 0.0 |
| GPT-3.5-turbo | 21 | 21 | 21.0 | 0.0 |
| GPT-4 | 29 | 30 | 29.7 | 0.5 |

Table 3: The summary statistics for the ToM tests responses.

On average, human participants achieved the highest total mean score of 31.3, while the GPT-3.5-turbo model had the lowest total score of 21.0. Remarkably, the total score of the GPT-4

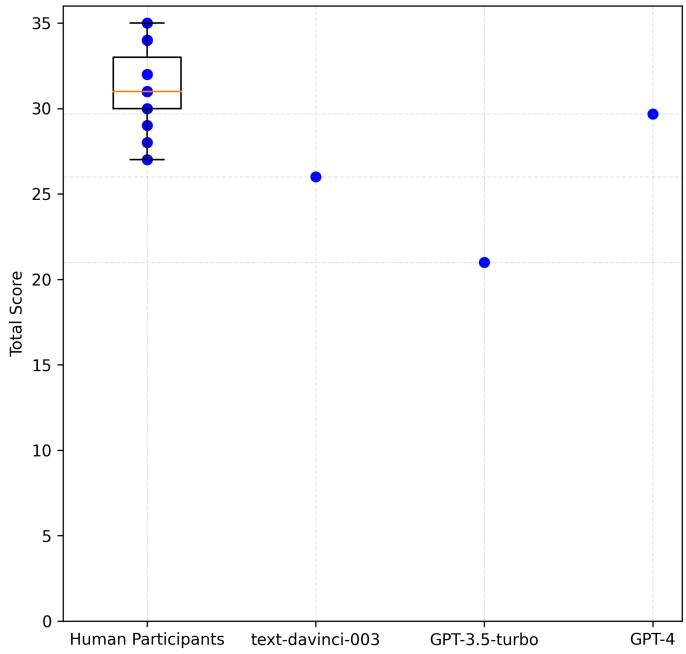


Figure 2: The summary statistics for the ToM tests responses.

model (29.7) closely resembles that of human participants (31.3). Moreover, all three GPT models maintain consistency, while human participants' answers exhibit a standard deviation of 2.2 with the only exception of the GPT-4 model, where only one response in three runs was different. Hence, we assume that the responses of each GPT model per each run are equal and further consider these to be a single measurement without a distribution. To evaluate the similarity between responses generated by GPT models and human participants, an independent-sample t-test was performed. The only significant p-value of 0.0023 was obtained for the GPT-3.5-turbo model given $\alpha = 0.05$, indicating that the model's performance differs from the human participants, not simply by chance.

## 4.2 Performance on ToM Stories

As mentioned in Subsection 2.1, the performance of GPT models compared to the human benchmark can be evaluated utilizing various types of ToM stories. Figures 3 and 4 illustrate the performance of human participants and three GPT models across different types of ToM stories.

To investigate whether the responses from the GPT models are from the same distribution as the responses from human participants, we have conducted a two-sample Kolmogorov-Smirnov test, which compares the distributions of two independent samples. The p-values of various GPT models per each ToM story type can be found in Table 4.

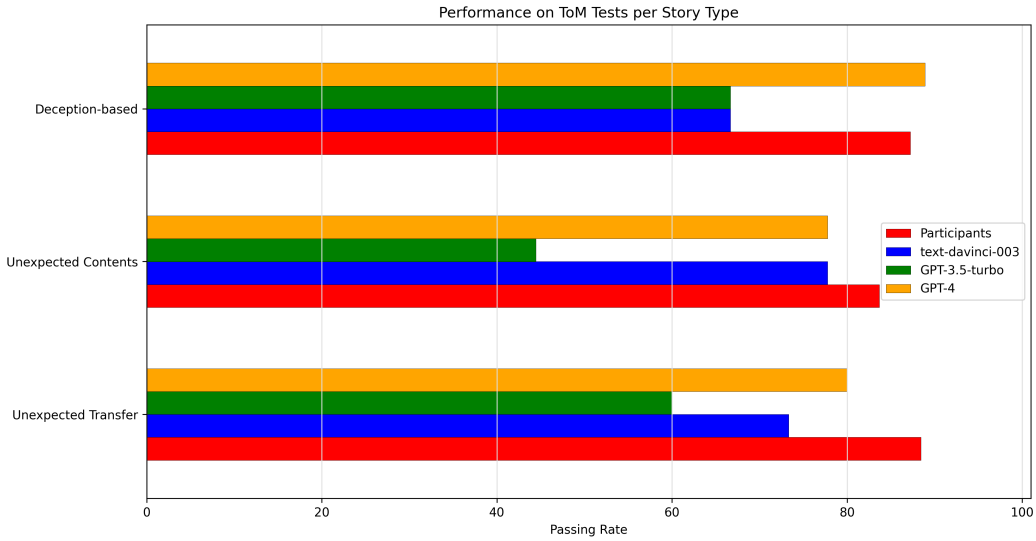


Figure 3: The performance of human participants and three GPT models across different types of ToM stories. On the y-axis, different types of ToM stories are displayed. On the x-axis, the average passing rate is depicted. Each horizontal bar represents either the human participants' performance or the performance of one of the GPT models based on the bar's color.

| GPT model | Unexpected Transfer | Unexpected Contents | Deception-based |
|---|---|---|---|
| text-davinci-003 | 0.678 | 0.989 | 0.536 |
| GPT-3.5-turbo | 0.184 | 0.126 | 0.536 |
| GPT-4 | 0.678 | 0.989 | 0.1 |

Table 4: The p-values of various GPT models per ToM story type.

The null hypothesis cannot be rejected for any GPT model for an assumed value of $\alpha = 0.05$, concluding that the responses provided by GPT models might be (but not necessarily are) from the same distribution as those provided by human participants.

From Figure 3, it can be seen that the GPT-4 model performs similarly to human participants on deception-based tasks (87% and 89% correspondingly). It performs worse than the human participants on unexpected transfer tasks (80% vs. 88%) and unexpected contents tasks (78% vs. 84%). The GPT-4 model outperforms the GPT-3.5-turbo model on all types of tasks. The GPT-4 model outperforms the text-davinci-003 model on the unexpected transfer (80% vs. 73%) and deception-based (92% vs. 67%) tasks, and it performs similarly to the text-davinci-003 on the unexpected contents tasks (78%).

The GPT-3.5-turbo model performs worse than human participants and the GPT-4 model on all tasks, with a passing rate of 60% on the unexpected transfer tasks, 44% on the unexpected contents tasks, and 67% on the deception-based tasks. The GPT-3.5-turbo model performs worse than the

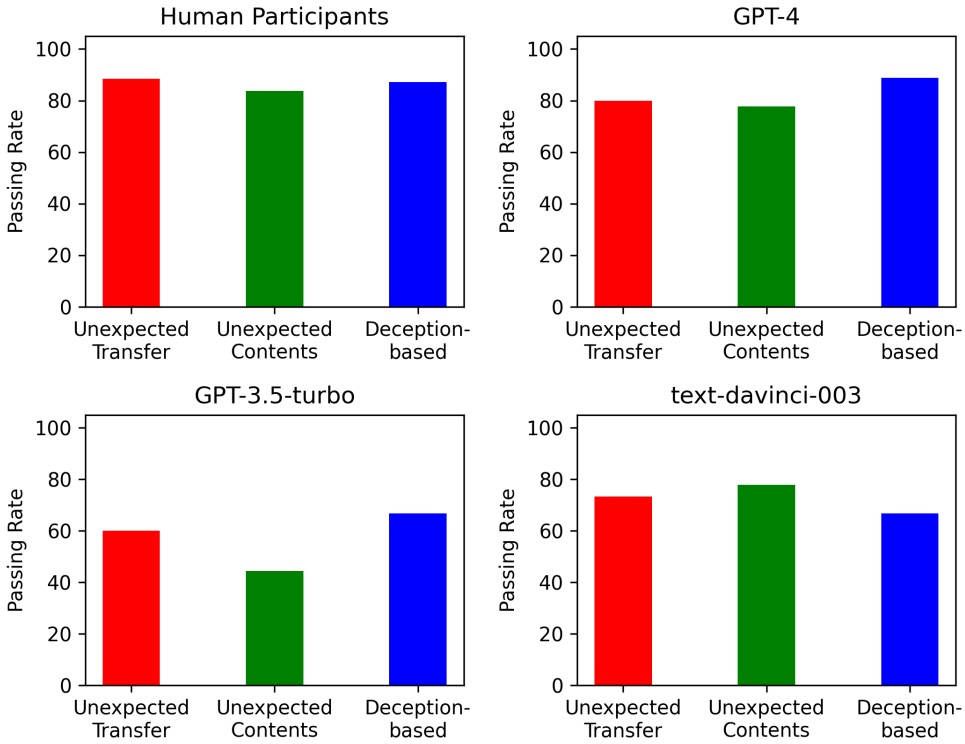


Figure 4: The performance of human participants and three GPT models across different types of ToM stories. On the x-axis, different types of ToM stories are displayed. On the y-axis, the average passing rate is depicted.

text-davinci-003 model on the unexpected transfer (60% vs. 74%) and unexpected contents tasks (44% vs. 78%) and alike on the deception-based tasks.

The text-davinci-003 model performs worse than human participants on all tasks, with a passing rate of 73% on the unexpected transfer tasks, 78% on the unexpected contents tasks, and 67% on the deception-based tasks. The text-davinci model performs similarly to the GPT-4 model on the unexpected contents tasks and to the GPT-3.5-turbo model on the deception-based tasks. It performs worse than the GPT-4 model on the unexpected transfer tasks (73% vs. 80%) and deception-based tasks (67% vs. 89%). The text-davinci-003 model outperforms the GPT-3.5-turbo model on the unexpected transfer tasks (73% vs. 60%) and unexpected contents tasks (78% vs. 44%).

## 4.3 Performance on ToM Questions

As mentioned in Subsection 2.1, the performance of GPT models compared to the human benchmark can also be evaluated utilizing various categories of ToM questions. Figures 5 and 6

illustrate the performance of human participants and three GPT models across different types of ToM questions.

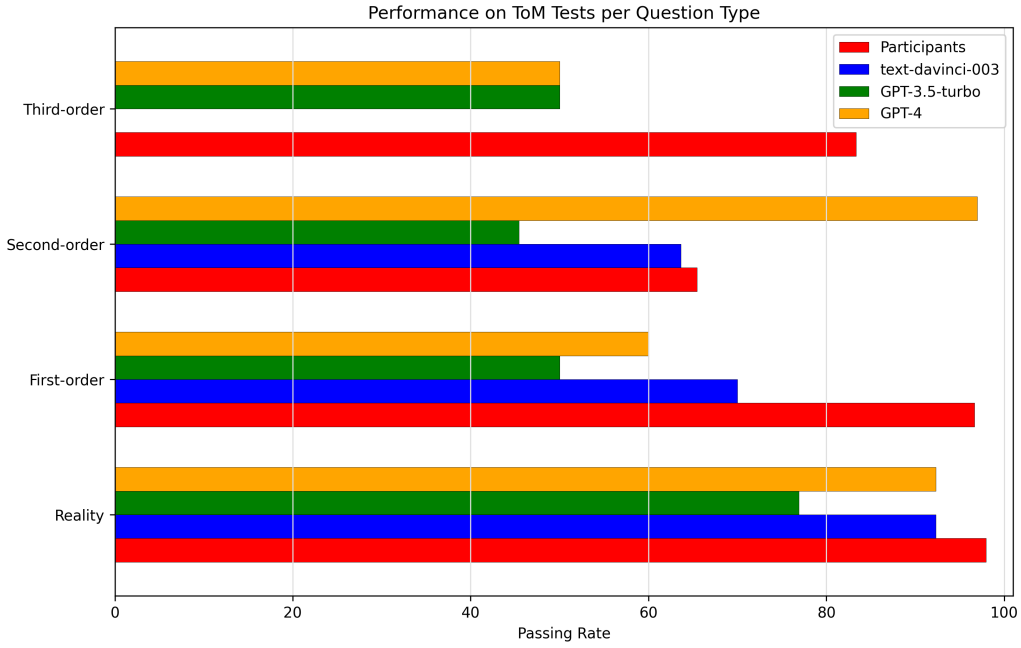


Figure 5: The performance of human participants and three GPT models across different types of ToM questions. On the y-axis, different types of ToM questions are displayed. On the x-axis, the average passing rate is depicted. Each horizontal bar represents either the human participants' performance or the performance of one of the GPT models based on the bar's color.

Similarly to the various types of ToM stories, we have conducted the two-sample Kolmogorov-Smirnov test for various types of ToM questions. The p-values of various GPT models per each ToM question type can be found in Table 5.

| GPT model | Reality | Firs-order | Second-order | Third-order |
|---|---|---|---|---|
| text-davinci-003 | 0.999 | 0.787 | 0.211 | 0.026 |
| GPT-3.5-turbo | 0.898 | 0.168 | 0.075 | 0.679 |
| GPT-4 | 0.999 | 0.418 | 0.004 | 0.679 |

Table 5: The p-values of various GPT models per ToM question type.

With a chosen significance level of $\alpha = 0.05$, we can reject the null hypothesis for the text-davinci-003 model on third-order false-belief questions and for the GPT-4 model on second-order false-belief questions. This indicates that the responses from these models on those specific question types do not follow the same distribution.

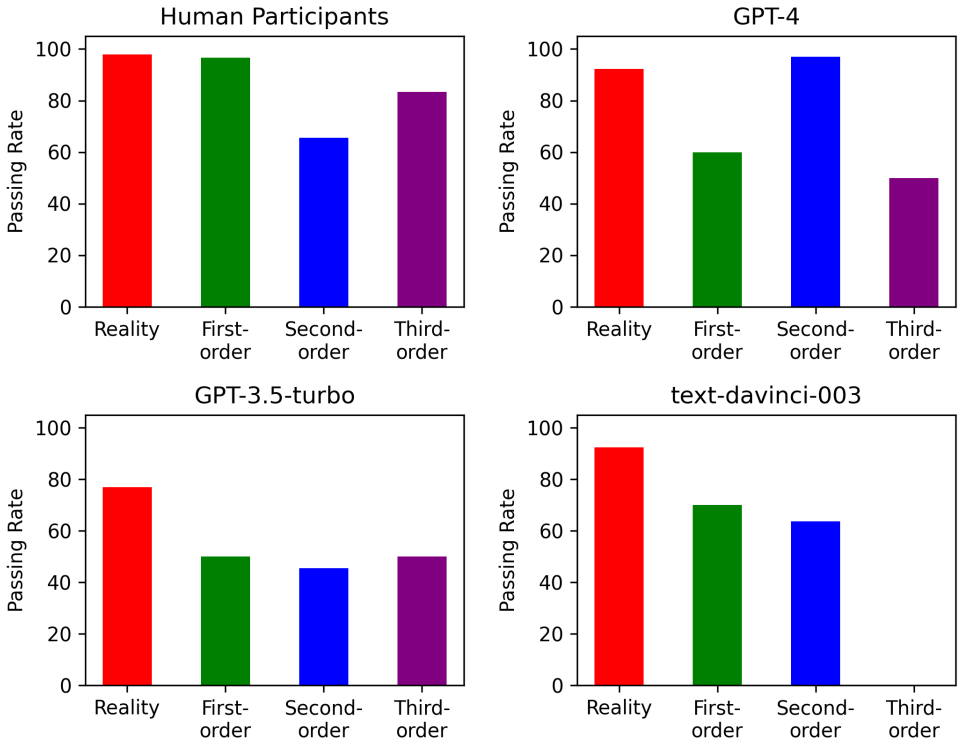


Figure 6: The performance of human participants and three GPT models across different types of ToM questions. On the x-axis, different types of ToM questions are displayed. On the y-axis, the average passing rate is depicted.

From Figure 5, it can be seen that the performance of the GPT-4 model surpasses that of human participants only on the second-order false-belief questions, achieving the passing rate of 97% compared to the humans' 65%. However, for the other types of ToM questions, GPT-4 performs worse than human participants. It achieves a passing rate of 92% on the reality questions (vs. 98%), 60% on the first-order false-belief questions (vs. 97%), and 50% on the third-order false-belief questions (vs. 83%). Compared to other models, GPT-4 outperforms all of them on the second-order false-belief questions. It performs similarly to the GPT-3.5-turbo model on the third-order false-belief questions and to the text-davinci-003 model on the reality questions. However, GPT-4 performs worse than the text-davinci-003 model on the first-order false-belief questions (60% vs. 70%).

The GPT-3.5-turbo model falls short of human participants' performance, scoring 77% on the reality questions, 50% on the first-order false-belief questions, 45% on the second-order false-belief questions, and 50% on the third-order false-belief questions. Compared to other GPT models, GPT-3.5-turbo performs better than the text-davinci-003 model (50% vs. 0%) and similarly to the GPT-4 model on third-order false-belief questions. On all other questions, the GPT-3.5-turbo

model performs worse than other models.

The text-davinci-003 model performs similarly to human participants on the second-order false-belief questions (64% and 65%) but worse on the other question types. It scores 93% on the reality questions, 70% on the first-order false-belief questions, and 0% on the third-order false-belief questions. Compared to other GPT models, text-davinci-003 performs better on the first-order false-belief questions and worse on the third-order false-belief questions. Regarding the reality questions, text-davinci-003 performs similarly to the GPT-4 model and outperforms the GPT-3.5-turbo model (92% vs. 77%). For the second-order false-belief questions, text-davinci-003 performs better than the GPT-3.5-turbo model (64% vs. 45%) and worse than the GPT-4 model (64% vs. 100%).

# 5 Discussion and Conclusion

This section focuses on the analysis of the results obtained in Section 4, as well as the discussion of the limitations encountered during the research. Furthermore, it draws conclusions based on the findings and offers insights into potential areas for future research.

## 5.1 Discussion

Human participants demonstrated outstanding performance on the reality and first-order false-belief questions and somewhat lower performance on the third-order false-belief questions. Noticeably, their performance declines considerably when it comes to the second-order false-belief questions, with the GPT-4 model significantly outperforming human participants. Upon closer examination of the second-order false-belief tests and discussing the participants' responses, we identified a potential issue related to the sequence and length of ToM questions. For instance, in the case of FBT_10, the first question presented is "What is inside the pack?" which is a reality question. Following it, the second-order false-belief question: "Tim is asked what he thinks another child, who hasn't seen the pack's content before, would guess is inside the pack. Tim answers, 'M&M's.' Is it a correct answer?" Due to the proximity of these questions in the ToM_tests.CSV file and the participants only paying attention to the following part of the question, "Tim is asked what he thinks ... is inside the pack," many participants have incorrectly answered "No, it is not" for this second-order false-belief question. The initial instruction clearly stated that there is no connection between the questions and that each question should be answered solely based on the information provided within the story itself and the corresponding question. Thus, human participants should allocate more attention and read every question more carefully.

Turning our attention to GPTs, it can be seen that the GPT-4 model exhibits a passing rate that closely aligns with that of human participants on most types of ToM tests and surpasses it in the second-order false-belief tests, achieving a passing rate of 97%. Therefore, we can infer that GPT-4 closely resembles human performance in ToM tests. The GPT-3.5-turbo model generally performs worse than human participants, GPT-4, and text-davinci-003 in all ToM tests, except the third-order false-belief tests, where it outperforms the text-davinci-003 model. This discrepancy may be attributed to the fact that the GPT-3.5-turbo model is specifically fine-tuned as a fast-performance dialogue model optimized for speed. As a result, it performs well in producing coherent responses, although they may not always be logical and reasonable within the given context. Another intriguing observation concerns the performance of the text-davinci-003 model on the third-order

false-belief tests. The model did not pass any tests in this category, indicating its inability to grasp the concept of higher-order false beliefs. This limitation becomes evident when examining Figure 5, as the model's performance consistently deteriorates with increasing false-belief order. One possible explanation for this observation is that the text-davinci-003 model is the oldest among the compared GPT models and, thus, might not be fine-tuned very well for higher-order false-belief tasks. It is important to note here that the baseline or human benchmark used in evaluating the performance of GPT models is derived from the responses of highly educated human adults, who can be expected to perform above average on such tests. Considering this, it is remarkable to observe the level of performance achieved by the current GPT models.

However, determining whether the answers provided by human participants and GPTs are similar in nature is challenging. The experiments reveal a standard deviation of 2.2 among human participants, which can be attributed to variations in ToM levels among individuals. In contrast, each run of a GPT model consistently produced identical responses, which we previously acknowledged as discrete measures rather than a distribution. To introduce some diversity in the responses, a temperature parameter can be utilized, which influences the randomness of the model's responses. Lowering the temperature value results in sharpening the probabilities of predicted words. This means that the most likely word is selected with a higher probability, leading to more conservative and predictable generated responses. In contrast, raising the temperature value flattens the predicted word probabilities, making all words more equally likely to be chosen. This promotes more creative and diverse response generation as the model becomes more inclined to produce unusual or unexpected words. Higher temperature encourages exploration, allowing for different responses in multiple runs, while lower temperature favors exploitation by generating more deterministic and consistent outputs.

Another crucial aspect is whether the presented tests can effectively assess ToM in LLMs. In humans, performance on such tests is known to be correlated with real-world social abilities, whereas in LLMs, it is currently an open question what high scores on ToM tests are predictive for. Moreover, general questions can be raised regarding the suitability of tests designed for humans when applied to LLMs [Hagendorff, 2023]. Lastly, the validity of ToM tests, in general, must be questioned. While in children, the inseparability of language and ToM competence is a major concern [Dörrenberg et al., 2018], language is less of an issue for LLMs. However, other underlying mechanisms involved when answering these ToM tests, such as memory or executive functioning [Launay et al., 2015], might have relevant parallels in LLMs. Exploring these parallels can be considered in future research.

In supporting the emergence of ToM in LLMs, we should discuss the concept of abstract representations. Just as humans exhibit some degree of abstraction in representing the world around them [Lynn et al., 2020], LLMs might also develop abstract representations by utilizing neural networks and deep learning. Each hidden layer in the network may represent an increasing complexity level of abstraction. This parallel suggests a similarity in how both humans and LLMs demonstrate the emergence of ToM. The custom-developed ToM tests had the same underlying principle as the baseline ToM tests and were never encountered by LLMs before, yet the models still performed well. This finding further supports the notion that both humans and LLMs may utilize a similar concept of abstract representations in their understanding and reasoning abilities, and hence LLMs might have acquired ToM-like abilities by the same means. Another similarity is the instruction tuning employed for adapting pre-trained LLMs for specific NLP tasks. As previously discussed, the parameters are fine-tuned through instruction tuning during the fine-tuning stage of

the training. Remarkably, this mirrors how humans learn during childhood, where instructions and feedback play a crucial role in refining their understanding and behavior.

## 5.2 Limitations

The first limitation of this bachelor project revolves around the developed ToM tests. As mentioned earlier, these tests are based on traditional false-belief tasks and share the same structure and underlying principles. However, as discussed in Subsection 2.2, LLMs excel in learning such structural patterns. Hence, the developed ToM tests may not be the most effective approach for evaluating ToM in LLMs. Therefore, other modifications of ToM tests should be considered. Alternatively, researchers could create entirely new ToM tests that deviate from the structure of existing ones. This approach would provide a more accurate evaluation of whether LLMs possess ToM since they would be encountering these novel tests for the first time, eliminating the potential bias of learned structural patterns.

Another limitation revolves around the relatively poor prompting of the GPT models in this project. Enhanced prompting would yield improved outcomes. As mentioned in [Moghaddam and Honey, 2023], appropriate prompting enhances ToM reasoning in LLMs, emphasizing the context-dependent nature of LLM cognitive capabilities. This improvement can be achieved by refining the instructions or modifying the type and content of each ToM question. For instance, employing fill-in-the-blank questions instead of the somewhat open-ended questions utilized in this bachelor project could lead to more precise LLMs' responses.

Finally, introducing an evaluation method can enhance and automate the evaluation process. As for now, a human rater has to grade the LLM responses, which takes time. By implementing such an evaluation method, it becomes possible to optimize and streamline the evaluation of LLM responses on ToM tests.

## 5.3 Conclusion and Further Research

This bachelor thesis aimed to investigate whether current LLMs exhibit emergent ToM. To achieve this, ToM tests were developed, and the responses of LLMs were compared to a human benchmark. The results indicated that recent LLMs such as GPT-4 closely resemble the performance of highly educated human participants on these tests.

As discussed in Subsection 2.3, there are two prevailing perspectives on emergent ToM in LLMs. On the one hand, researchers such as Kosinski argue that current LLMs may possess ToM as it emerges as a byproduct of increasingly improved language capabilities. On the other hand, researchers such as Ullman claim that LLMs might not possess ToM at all. Instead, they simply learn the patterns or structures of ToM tests. With sufficient training data, LLMs can learn these patterns and fine-tune their network weights to achieve high test scores. Yet, it should be pointed out that there is at least a small degree to which a parallel argument can be made for humans.

Based on the findings obtained, we suggest that current LLMs of the GPT family achieve a level of performance on both standardized and novel ToM tasks comparable to that of highly-educated participants. Further research will be needed to assess the robustness of this result, for example, by testing LLMs on a broader range of tasks relevant to ToM. Other prospects include investigating the impact of temperature or prompting variations on the GPTs' responses.

# References


[Apperly and Butterfill, 2009] Apperly, I. A. and Butterfill, S. A. (2009). Do humans have two systems to track beliefs and belief-like states? *Psychological review*, 116(4):953.

[Bahdanau et al., 2014] Bahdanau, D., Cho, K., and Bengio, Y. (2014). Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*.

[Baron-Cohen, 1991] Baron-Cohen, S. (1991). Precursors to a theory of mind: Understanding attention in others. *Natural theories of mind: Evolution, development and simulation of everyday mindreading*, 1:233–251.

[Baron-Cohen et al., 1985] Baron-Cohen, S., Leslie, A. M., and Frith, U. (1985). Does the autistic child have a "theory of mind"? *Cognition*, 21(1):37–46.

[Bowman, 2023] Bowman, S. R. (2023). Eight things to know about large language models. *arXiv preprint arXiv:2304.00612*.

[Brown et al., 2020] Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al. (2020). Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

[Cohen, 2021] Cohen, M. (2021). Exploring roberta's theory of mind through textual entailment. *Philarchive*.

[Dennett, 1978] Dennett, D. C. (1978). Beliefs about beliefs [p&w, sr&b]. *Behavioral and Brain sciences*, 1(4):568–570.

[Dennett, 1983] Dennett, D. C. (1983). Intentional systems in cognitive ethology: The "panglossian paradigm" defended. *Behavioral and Brain Sciences*, 6(3):343–355.

[Dörrenberg et al., 2018] Dörrenberg, S., Rakoczy, H., and Liszkowski, U. (2018). How (not) to measure infant theory of mind: Testing the replicability and validity of four non-verbal measures. *Cognitive Development*, 46:12–30.

[Duijn, 2016] Duijn, M. J. v. (2016). The lazy mindreader. a humanities perspective on mindreading and multiple-order intentionality. *Netherlands: Koninklijke Wöhrman*.

[Esteva et al., 2017] Esteva, A., Kuprel, B., Novoa, R. A., Ko, J., Swetter, S. M., Blau, H. M., and Thrun, S. (2017). Dermatologist-level classification of skin cancer with deep neural networks. *nature*, 542(7639):115–118.

[Frith and Frith, 2005] Frith, C. and Frith, U. (2005). Theory of mind. *Current biology*, 15(17):R644–R645.

[Hagendorff, 2023] Hagendorff, T. (2023). Machine psychology: Investigating emergent capabilities and behavior in large language models using psychological methods. *arXiv preprint arXiv:2303.13988*.

[Kinderman et al., 1998] Kinderman, P., Dunbar, R., and Bentall, R. P. (1998). Theory-of-mind deficits and causal attributions. *British journal of Psychology*, 89(2):191–204.

[Kosinski, 2023] Kosinski, M. (2023). Theory of mind may have spontaneously emerged in large language models. *arXiv preprint arXiv:2302.02083*.

[Launay et al., 2015] Launay, J., Pearce, E., Wlodarski, R., van Duijn, M., Carney, J., and Dunbar, R. I. (2015). Higher-order mentalising and executive functioning. *Personality and individual differences*, 86:6–14.

[Liddle and Nettle, 2006] Liddle, B. and Nettle, D. (2006). Higher-order theory of mind and social competence in school-age children. *Journal of Cultural and Evolutionary Psychology*, 4(3-4):231–244.

[Lynn et al., 2020] Lynn, C. W., Kahn, A. E., Nyema, N., and Bassett, D. S. (2020). Abstract representations of events arise from mental errors in learning and memory. *Nature communications*, 11(1):2313.

[Manning, 2022] Manning, C. D. (2022). Human language understanding & reasoning. *Daedalus*, 151(2):127–138.

[Meltzoff, 1999] Meltzoff, A. N. (1999). Origins of theory of mind, cognition and communication. *Journal of communication disorders*, 32(4):251–269.

[Meltzoff and Decety, 2003] Meltzoff, A. N. and Decety, J. (2003). What imitation tells us about social cognition: a rapprochement between developmental psychology and cognitive neuroscience. *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences*, 358(1431):491–500.

[Milligan et al., 2007] Milligan, K., Astington, J. W., and Dack, L. A. (2007). Language and theory of mind: Meta-analysis of the relation between language ability and false-belief understanding. *Child development*, 78(2):622–646.

[Moghaddam and Honey, 2023] Moghaddam, S. R. and Honey, C. J. (2023). Boosting theory-of-mind performance in large language models via prompting. *arXiv preprint arXiv:2304.11490*.

[Nasr et al., 2019] Nasr, K., Viswanathan, P., and Nieder, A. (2019). Number detectors spontaneously emerge in a deep neural network designed for visual object recognition. *Science advances*, 5(5):eaav7903.

[O'Brien et al., 2011] O'Brien, M., Miner Weaver, J., Nelson, J. A., Calkins, S. D., Leerkes, E. M., and Marcovitch, S. (2011). Longitudinal associations between children's understanding of emotions and theory of mind. *Cognition & emotion*, 25(6):1074–1086.

[Perner et al., 1987] Perner, J., Leekam, S. R., and Wimmer, H. (1987). Three-year-olds' difficulty with false belief: The case for a conceptual deficit. *British journal of developmental psychology*, 5(2):125–137.

[Pyers and Senghas, 2009] Pyers, J. E. and Senghas, A. (2009). Language promotes false-belief understanding: Evidence from learners of a new sign language. *Psychological science*, 20(7):805–812.

[Radford et al., 2018] Radford, A., Narasimhan, K., Salimans, T., Sutskever, I., et al. (2018). Improving language understanding by generative pre-training.

[Sap et al., 2022] Sap, M., LeBras, R., Fried, D., and Choi, Y. (2022). Neural theory-of-mind? on the limits of social intelligence in large lms. *arXiv preprint arXiv:2210.13312*.

[Saxe and Kanwisher, 2003] Saxe, R. and Kanwisher, N. (2003). People thinking about thinking people: the role of the temporo-parietal junction in "theory of mind". *Neuroimage*, 19(4):1835–1842.

[Silver et al., 2016] Silver, D., Huang, A., Maddison, C. J., Guez, A., Sifre, L., Van Den Driessche, G., Schrittwieser, J., Antonoglou, I., Panneershelvam, V., Lanctot, M., et al. (2016). Mastering the game of go with deep neural networks and tree search. *nature*, 529(7587):484–489.

[Stoianov and Zorzi, 2012] Stoianov, I. and Zorzi, M. (2012). Emergence of a'visual number sense'in hierarchical generative models. *Nature neuroscience*, 15(2):194–196.

[Ullman, 2023] Ullman, T. (2023). Large language models fail on trivial alterations to theory-of-mind tasks. *arXiv preprint arXiv:2302.08399*.

[Vaswani et al., 2017] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. (2017). Attention is all you need. *Advances in neural information processing systems*, 30.

[Westby and Robinson, 2014] Westby, C. and Robinson, L. (2014). A developmental perspective for promoting theory of mind. *Topics in language disorders*, 34(4):362–382.

[Wimmer and Perner, 1983] Wimmer, H. and Perner, J. (1983). Beliefs about beliefs: Representation and constraining function of wrong beliefs in young children's understanding of deception. *Cognition*, 13(1):103–128.

# A Project Files

The GPT UI, API, auxiliary functions, ToM tests, GPT and human participants' responses, and performance plots can be accessed through the GitHub repository link.

# B GPT UI

The GPT UI is developed in Python utilizing the Tkinter toolkit. The GPT UI project consists of three main files:

- GPT_UI.py contains the UI. We will be discussing the functionality of this file in the rest of this appendix;

- GPT_API.py contains three GPT APIs. Those are text-davinci-003, GPT-3.5-turbo, and GPT-4. Each model receives the prompt, temperature, and the maximum number of tokens. Based on the provided hyperparameter values, the response is then generated;
- Auxiliary.py contains various auxiliary functions for this bachelor project. The function *responses()* returns the responses of human participants and GPT models. The function *summary_statistics()* calculates the summary statistics and generates a related boxplot. Finally, the functions *performance_per_story_type()* and *performance_per_question_type()* generate performance plots based on a story or question type.

The GPT UI consists of four distinct sections: *Personal Key*, *Read ToM Test(s)*, *Run ToM Test(s)*, and *Save Results*. Initially, all sections except the *Personal Key* section are disabled, and each subsequent section becomes active only when specific conditions are met. For example, to unlock the *Read ToM Test(s)* section, the personal OpenAI key must be entered.

### Personal Key

In this section, users can insert their personal OpenAI key to utilize the GPT API. For security reasons, each character entered in the key box is replaced with an asterisk (*). To submit the key, the *OK* button can be pressed. To change the key, the *Change* button can be pressed. Users also have the option to use the default key by selecting the *Use default key* checkbox. The default key must be entered directly in the UI code under *PLEASE, SET THE DEFAULT KEY HERE*. Once the personal key is manually entered or the default key checkbox is selected, the system will perform a validation check to verify the authenticity of the key. If the key format is incorrect or the key itself is invalid, an appropriate error message such as "Unauthorized OpenAI key!" will be displayed. Once a valid key has been successfully entered, the *Read ToM Test(s)* section will become accessible.

### Read ToM Test(s)

In this section, users can either read ToM tests from a CSV file or enter a test manually. All tests will be stored in the pandas DataFrame.

Users can select one of the option buttons to choose their preferred method. To read ToM tests from a CSV file, *Read CSV file* option should be selected. To insert a ToM test manually, the *Insert test manually* option should be chosen. Please note that only one test can be manually entered into the system in this GPT UI version.

When one of the options is selected, the corresponding functionality will be unlocked. To read ToM tests from a CSV file, users can click the *Browse file* button, which opens a dialog box where a CSV file can be selected. Once the file is selected, a system message will inform the user whether the action was successful. The file's name will also be displayed above the *Show tests* button.

To manually enter a test, users need to fill in the relevant fields denoted by *ID*, *Description*, *Question*, and *Correct Answer*. It is worth mentioning that this UI does not validate the entered information, so it is the user's responsibility to ensure the test is entered correctly. If one of the fields is empty, the corresponding error message will be displayed, preventing the user from inserting the uncompleted test. Once each field has been correctly entered, users can press the *Add test* button to add the test into the system.

Finally, to view the read from a CSV file tests or a manually entered test, users can click one of the *Show test(s)* buttons. This action opens a new window where the test(s) can be observed.

Once the ToM tests are successfully inserted into the pandas DataFrame, the *Run ToM Test(s)* section will be unlocked.

### Run ToM Test(s)

In this section, users can conduct ToM experiments by prompting the GPT model with an instruction, a description of the test, and a question.

The *Instruction* field provides instructions to the GPT model and can be changed upon need. Next, the number of questions can be set. By default, this value is set to 3 for the ToM tests read from a CSV file and to 1 for a test entered manually. Following, three hyperparameter values can be adjusted. The first is the engine model denoted by *Choose the model*, which offers three options: *text-davinci-003*, *gpt-3.5-turbo*, and *gpt-4*. By default, the *gpt-3.5-turbo* is selected. The second hyperparameter is the temperature denoted by *Set the temperature*, a value between 0 and 1. It can be increased or decreased by 0.1 using the slider on the right. The default value is set to 0, corresponding to the responses with the highest probability. Finally, the third option is the maximum number of tokens denoted by *Set the max number of tokens*. It is important to note that tokens may not necessarily correspond to words and can include letters or combinations of letters like word roots. The default value is set to 30. Once users are satisfied with the hyperparameter values, they can click the *Run experiment* button. The parameter values, instruction, descriptions of the tests, and questions will be provided to the GPT API, and a response for each question will be generated. Upon completion, a system message will be displayed. To view the response(s) along with the test(s), users can press the *Show results* button.

Upon successful completion, the *Save results* section will be unlocked.

### Save Results

In this section, users can save the GPT responses in a CSV file with a specified name. By default, the file's name is set to ToM_results.csv. Users need to click the *Save* button to save the file. Upon successful saving, a system message will appear confirming the completion. The CSV file will be stored in the project's folder.

### Additional functionality

At any point, the GPT UI can be reset to its default state by clicking the *Reset* button. This action will clear all fields, revert the hyperparameter values to their default states, and delete all ToM tests from the pandas DataFrames. To cancel the GPT UI, the *Close* button can be pressed.

The *Performance plot* button generates a plot that showcases the averaged passing percentage per ToM story or question type. The generated plot can be chosen in the *performance_plot()* function. This plot illustrates the performance of both the GPT LLMs and human participants. However, due to the limitations of evaluating the test results, the function associated with this button is not directly linked to the results obtained. Instead, the parameters for generating this plot, e.g., the participants and GPT models responses should be manually entered in the *responses()* function in the Auxiliary.py file. These limitations arise from the fact that the GPT model can provide multiple correct answers to the same question, and the primary goal of this project is not to develop

a system that grades these responses. Please also note that the *Performance plot* button will always be active.

# C ToM Tests

## Structure

The ToM stories among the questions can be found in the ToM_tests.csv file. The structure of the ToM_tests.csv file is as follows:

- *ID* – the unique identifier for the ToM story, e.g., FBT_10;
- *Description* – the content of the story, e.g.,

  *Mike and Roy are conducting an experiment at primary school for the cognitive psychology course. They are now sitting in the classroom with Tim, one of the pupils. They show Tim a pack that looks like a pack of M&Ms. However, unknown to any participant, the pack actually contains Skittles. Tim is then asked what he thinks is inside the pack, and he undoubtedly answers, "M&M's." After that, Mike and Roy open the pack, and Tim sees the content of the pack. The pack is then closed.*

- *Question 1*, *Question 2* and *Question 3* – the questions related to the story. Each question has a unique identifier based on the story's ID, e.g.,

  FBT_9.1: Which popcorn did Mateo buy for himself?
  FBT_9.2: Did Mateo tell the truth that he likes his popcorn?
  FBT_9.3: Does Emily believe that Mateo likes his popcorn?

  Please note the question identifiers are not depicted in the ToM_tests.csv file.

- *Correct Answer 1*, *Correct Answer 2* and *Correct Answer 3* – the correct answers corresponding to the respective questions, e.g.,

  1. Skittles.
  2. Yes, it is.
  3. No, it isn't.

## ToM Tests

### FBT_1

*Sander and Emma are sister and brother. They are playing outside in the small garden. Their father comes and gives Emma a small pack of gummy bears but not Sander because he didn't behave well lately. Emma eats some gummy bears and decides to eat the remainder later. She puts the pack with candies on the chair, which is located under the tree at the farthest corner of the garden. She doesn't give any gummy bears to Sander. After that, Emma goes inside the house to drink some water. Sander is now alone in the garden. He is upset because Emma didn't share the candies with him. Sander decides to steal the gummy bears. He takes the pack with sweets from the chair and hides it in the bush nearby. At the same time, Emma*

*looks out of the window and sees Sander replacing the gummy bears. However, Sander does not notice that Emma sees him. Subsequently, Emma comes back to the garden and tells Sander she wants to eat some gummy bears.*

**FBT_1.1**: Where are the gummy bears? *In the bush.*
**FBT_1.2**: Where will Emma look for the gummy bears? *In the bush.*
**FBT_1.3**: Where does Sander think Emma will look for the gummy bears? *On the chair.*

**FBT_2**

*Oliver and Charlotte are playing in the room with some toys. Peter, a friend of Oliver, comes by and asks Oliver if he can borrow his bicycle. Oliver says yes. Oliver and Peter then walk outside the house, where Oliver hands the bicycle to Peter and receives the chocolate bar in return for his favor. When Oliver is back in the room, he teases Charlotte because he has a chocolate bar, but she doesn't. Charlotte is not happy about that. Oliver puts the chocolate bar in his backpack to eat it later. Peter then returns to the room and tells Oliver that there is something wrong with the bicycle and that he needs help. Both boys leave the room again, and Charlotte is now alone. She is sad because Oliver was mean to her. Charlotte decides to hide the candy in a different place so that Oliver will not find it later. She grabs the chocolate bar from the backpack and puts it under the bed. At the same time, Oliver finds out that to resolve the problem with the bicycle, he needs to grab some tools from his room, so he walks back. When he approaches the door of his room, he sees that Charlotte is hiding his chocolate bar under the bed, but Charlotte does not see him. Moments later, Oliver walks into the room, takes the tools, and returns to his friend. Now the bicycle is fixed, and Oliver is back in the room. Oliver tells Charlotte he wants to eat his chocolate bar.*

**FBT_2.1**: Where is the chocolate bar? *Under the bed.*
**FBT_2.2**: Where will Oliver look for the chocolate bar? *Under the bed.*
**FBT_2.3**: Where does Charlotte think Oliver will look for the chocolate bar? *In his backpack.*

**FBT_3**

*John and Hannah have three children: Liam, Thomas, and Ivy. The children are in the summer school. They are having a math quiz now. Thomas wins the quiz with the highest score, while Liam is second and Ivy is fifth. Because of his win, Thomas receives a present from the teacher - a robot. Thomas is pleased and can't wait to play with it. He puts the robot in the drawer of his desk. Liam is sorrowful. He always loses to his brother when it comes to the quiz, and he also wanted to have a robot like that for a long time. It's break time! Thomas decides to go to the canteen to grab a snack. Liam sees a chance to play with the robot while Thomas is away. He takes the robot from the drawer and goes to the play corner. He likes the robot, and he wishes he could have it. While Liam is playing with the robot, Thomas passes by the classroom and sees it. He is slightly surprised, but it is not a big deal for him as Thomas knows his brother well and is happy to share his toys. Thomas then sees his friend, Mark, in the hall and walks towards him to chat a bit. Meanwhile, Liam returns the robot to the drawer because he respects and loves his brother. Thomas does not see it. Liam leaves the classroom as well for a short walk. Ivy, however, thinks that this is all very unfair to Liam. She likes her brother Liam more than Thomas and decides to take the robot and give it to Liam later as a present. Ivy puts the robot in her backpack without anyone seeing it. Then*

*Thomas and Liam return to the classroom, and Thomas would like to play with his robot. Thomas opens the drawer, but there is no robot.*

**FBT_3.1**: Where is the robot? *In Ivy's backpack.*
**FBT_3.2**: Who took Thomas' robot away? *Ivy.*
**FBT_3.3**: Who does Thomas think took his robot away? *Liam.*

**FBT_4**

*Max would like to become a pilot. It has been his dream since he was a child. Max is now ready to apply for flight school, but before he can do that, he needs to obtain a health certificate to allow him to fly commercial airliners. Max asks his friend Finn, who is already a student at that flight academy, where he can get that medical certificate. However, it is the first of April, and Finn decides to play a small prank. He gives Max the old address of the medical clinic. Max goes there and finds a grocery shop. When he asks the cashier, Isabella, whether he came to the correct address, she tells him the clinic he is looking for moved a few months ago to the new location. Max goes there. Fortunately, this time the address is correct. There, Max meets his other friend, Tim, who is also a friend of Finn. Max tells Tim what just happened, and to Max's surprise, Tim says that Finn has just pranked him too! So, Max decides to pull a prank on Finn as well. When Max and Finn meet later, Finn asks how it went with the medical certificate, and Max replies that it went very well. Max says that he found the medical clinic at the address given by Finn without any problems.*

**FBT_4.1**: Why did Finn prank Max? *Because it was April Fools' Day.*
**FBT_4.2**: Was Max honest with Finn that he found the clinic without any problem? *No, he wasn't.*
**FBT_4.3**: Does Finn think that Max knows that Finn pranked him about the location of the medical clinic? *Yes, he does.*

**FBT_5**

*Benjamin, Olivia, Felix, and William are roommates. They share a small apartment in Amsterdam with a fantastic view of the Prinsengracht. One evening Benjamin and Felix were hanging out in the living room, and Felix told Benjamin that Olivia was preparing a surprise party for Benjamin's birthday. Benjamin was somewhat surprised as he and Olivia just broke up and had a tough time. Out of curiosity, Benjamin asked Felix how he knew about the party, and Felix replied that William had told him last Friday when they were having drinks in the city. William said that Olivia told him she felt sorry because of the breakup with Benjamin and decided to organize a surprise party to show Benjamin that she still liked him. Benjamin was skeptical about everything Felix had just said. Still, Felix assured Benjamin that he told the truth, as Olivia is very good at keeping secrets. The boys then decided to go out and spent the whole night in the city center. The next day, Benjamin wondered whether Felix was telling the truth about the surprise party. Benjamin suspected that Felix had just pranked him. Later that day, Benjamin met Olivia in the living room and decided to ask about the party. Olivia was surprised that William had told Felix about the party. Olivia said she did not plan any party and never talked to Felix or William about it. Benjamin was confused. He started thinking Felix was either mistaken or lying and asked Olivia what she thought. Olivia replied that Felix is an honest person and she never caught him on a lie before. "However,"*

*added Olivia, "William might have pranked Felix as he had done many times before."*

**FBT_5.1**: According to Olivia, why did William tell Felix about the surprise party? *As a prank.*
**FBT_5.2**: What is more likely: Felix is mistaken about the surprise party, or he is just lying to Benjamin? *Felix is mistaken about the surprise party.*
**FBT_5.3**: Does Olivia think that Felix beliefs that she told William about the surprise party? *Yes, she does.*

**FBT_6**

*It's sunny today, and Liam and Oliver decided to go to Highland Park to enjoy some nice weather. After walking for some time, they see a hot dog truck. Liam asks the hot dog man, Jim, if they can have two hot dogs. Jim kindly replies: "Yes, of course, but it is only possible to pay by cash." Unfortunately, both Oliver and Liam only have credit cards with them. They ask Jim whether he will remain selling hot dogs here, to which Jim replies that he will stay here for a few hours. Oliver decides to quickly run to his house to get some cash, as he lives nearby, while Liam will stay in the park. Moments later, while waiting for Oliver, Liam sees Jim closing the hot dog truck. Liam asks Jim whether he is leaving. Jim replies that no one is buying hot dogs, so he is moving to North Square. Jim leaves the park in his truck. At the same time, Oliver arrived home and found some cash. He is about to leave the house when he sees Jim out of the window in his hot dog truck. He asks Jim: "I thought you said you would stay at Highland Park, to which Jim replies: "Yes, I did, but I couldn't sell any hot dogs, so I decided to change the location. I'm going to North Square now." Moments later, Liam calls Oliver and tells him he needs to work on the math assignment. So they decide to meet later. They did not talk about the hot dog truck. While studying, Liam realizes he needs Oliver's help. So, Liam is now on his way to Oliver's. When there, he rings the doorbell. Oliver's father opens the door. Liam asks whether Oliver is home to what the father replies: "He's just left. He said he is going to buy a hot dog from Jim."*

**FBT_6.1**: Where is Jim? *At North Square.*
**FBT_6.2**: Does Liam know that Oliver talked to Jim? *No, he doesn't.*
**FBT_6.3**: Where does Liam think Oliver has gone? *At Highland Park.*

**FBT_7**

*Mateo and Jack are at university today. After the classes, they walk to the bicycle parking to grab their bikes and cycle to Jack's home to work on a history project. At the parking spot, the boys notice that Jack's bicycle has a flat tire. They decide to go to Mateo's house first, as he lives closer, grab some tools and then go back and fix Jack's bike. To get to Mateo's place, they need to take a bus. At the bus stop, Jack and Mateo see their classmate Otis walking towards the car parking. Mateo and Jack ask Otis whether he can give them a lift to Mateo's house, and Otis agrees. However, Mateo realizes he left the project materials in the classroom, so he has to run back to school to grab the papers. They decide that Otis will pick Mateo and Jack up at the bus stop, and Mateo leaves them. After Mateo has left, Otis comes up with a better idea: they can just put Jack's bicycle in Otis's car trunk and bring it to Mateo's home. "Such a great idea," says Jack. Otis suggests he will pick up Jack and Mateo at the bicycle parking instead, and Jack agrees. Otis is now on his way to the car parking to get his*

*car while Jack is walking towards the bicycle parking. On the way, Otis meets Mateo and tells him he will pick him and Jack up at the bicycle parking and that Jack also knows about it. Moments later, Mateo reaches the school, finds the papers in the classroom, and is now on his way to the bicycle parking. Meanwhile, Jack realizes that Mateo will probably walk back to the bus stop, so Jack decides to go to school to inform Mateo that they're meeting at the bicycle parking. At school, Jack cannot find Mateo.*

**FBT_7.1**: Where will Otis pick up Jack and Mateo? *At the bicycle parking.*
**FBT_7.2**: Where does Mateo think Jack has gone? *The bicycle parking.*
**FBT_7.3**: Where does Jack think Mateo has gone? *The bus stop.*

**FBT_8**

*Henry and Chloe walk into the canteen to have lunch together. Henry orders a sandwich, yogurt, a plate of fresh fruits, and an apple juice. Chloe orders pasta and some fresh vegetables. She also wants to order an apple juice, but she always does so and decides to go for an orange juice this time. Then they walk to the table and sit down. Chloe says she forgot to grab a fork and walks back to the cash desk to take one. Meanwhile, Harry decides he doesn't want the apple juice and would like the orange juice instead. So he swaps the juices. When Chloe is back, she wonders whether she ordered the apple juice. Henry says: "Yes, you did. You always order the apple juice." To which Chloe says: "I'm quite sure I bought the orange juice."*

**FBT_8.1**: Which juice did Chloe buy? *The orange juice.*
**FBT_8.2**: Did Henry tell the truth about the juice? *No, he didn't.*
**FBT_8.3**: Does Chloe believe Henry told the truth? *No, she doesn't.*

**FBT_9**

*Mateo and Emily have a date at the cinema. There, they decide to buy some popcorn. There are two options: salty and sweet. Mateo doesn't like salty popcorn, but Emily doesn't know it. There is only one sweet popcorn left. Mateo asks Emily: "Which popcorn do you like?" and Emily replies: "I love the sweet one." Mateo buys the sweet popcorn for Emily and the salty one for himself. During the movie, Emily asks Mateo: "How do you like your popcorn?" And Mateo answers: "I like it a lot!" to what Emily says: "But you have barely touched it."*

**FBT_9.1**: Which popcorn did Mateo buy for himself? *The salty popcorn.*
**FBT_9.2**: Did Mateo tell the truth that he likes his popcorn? *No, he didn't.*
**FBT_9.3**: Does Emily believe that Mateo likes his popcorn? *No, she doesn't.*

**FBT_10**

*Mike and Roy are conducting an experiment at primary school for the cognitive psychology course. They are now sitting in the classroom with Tim, one of the pupils. They show Tim a pack that looks like a pack of M&Ms. However, unknown to any participant, the pack actually contains Skittles. Tim is then asked what he thinks is inside the pack, and he undoubtedly answers, "M&M's." After that, Mike and Roy open the pack, and Tim sees the content of the pack. The pack is then closed.*

**FBT_10.1**: What is inside the pack? *Skittles.*
**FBT_10.2**: Tim is asked what he thinks another child, who hasn't seen the pack's content

before, would guess is inside the pack. Tim answers, "M&M's." Is it a correct answer? *Yes, it is.*

**FBT_10.3**: Tim is asked what he thinks another child, who hasn't seen the pack's content before, would guess is inside the pack. Tim answers, "Skittles." Is it a correct answer? *No, it isn't.*

**FBT_11**

*There are two black plastic bags in the kitchen: one labeled as "potatoes" and the other as "tomatoes." However, unknown to anyone in the household, the labels have been wrongly placed as the bag labeled "potatoes" actually contains tomatoes, and the bag labeled "tomatoes" contains potatoes. Olga decides to cook some potatoes. She looks at the bags and reads the labels. Olga takes the one labeled "potatoes."*

**FBT_11.1**: What does Olga believe is inside the plastic bag she took? *Potatoes.*

**FBT_11.2**: Olga opens the plastic bag. What will she see inside? *Tomatoes.*

**FBT_11.3**: Olga opens the plastic bag and sees no potatoes but only tomatoes. She checks the other bag and realizes the problem. So, she swaps the labels on the plastic bags. Later, Inga, Olga's sister, decides to make a tomato salad. She takes the plastic bag labeled "tomatoes" and opens it. What will Inga see inside the plastic bag? *Tomatoes.*

**FBT_12**

*Sarah has five different notebooks labeled "Math," "French," "Biology," "Chemistry," and "Physics," which contain the notes for related classes. However, when writing those labels, Sarah made a couple of mistakes. The notebook labeled "Biology" contains the chemistry class notes. The notebook labeled "Chemistry" contains the notes from the physics class. Finally, the notebook labeled "Physics" contains the notes from the biology class. Sarah and Julia are studying in Sarah's living room. Julia would like to see Sarah's notes from the biology class. Sarah tells Julia that she can take the notebook from her room. Julia is now in Sarah's room looking at five notebooks. Julia doesn't know that the labels are wrong. She reads the labels.*

**FBT_12.1**: Notebook with which label will Julia take? *The notebook labeled "Biology."*

**FBT_12.2**: Julia opens the notebook labeled "Biology." What will Julia see inside this notebook? *Chemistry class notes.*

**FBT_12.3**: Which notebook should Julia take to have the physics class notes? *The notebook labeled "Chemistry."*