

Theory of mind performance of large language models: A comparative analysis of Turkish and English

Burcu Ünlütürk^{*}, Onur Bal

Department of Psychology, Yeditepe University, 26 Ağustos Yerleşkesi, Ataşehir İnönü Mah. Kayışdağı Cad, İstanbul, Turkey

ARTICLE INFO

Keywords:

Theory of mind
Large language models
First-order false belief
Second-order false belief

ABSTRACT

Theory of mind (ToM), understanding others' mental states, is a defining skill belonging to humans. Research assessing LLMs' ToM performance yields conflicting findings and leads to discussions about whether and how they could show ToM understanding. Psychological research indicates that the characteristics of a specific language can influence how mental states are represented and communicated. Thus, it is reasonable to expect language characteristics to influence how LLMs communicate with humans, especially when the conversation involves references to mental states. This study examines how these characteristics affect LLMs' ToM performance by evaluating GPT 3.5 and 4 performances in English and Turkish. Turkish provides an excellent contrast to English since Turkish has a different syntactic structure and special verbs, *san-* and *zannet-*, meaning "falsely believe." Using Open AI's Chat Completion API, we collected responses from GPT models for first- and second-order ToM scenarios in English and Turkish. Our innovative approach combined completion prompts and open-ended questions within the same chat session, offering deep insights into models' reasoning processes. Our data showed that while GPT models can respond accurately to standard ToM tasks (100% accuracy), their performance deteriorates (below chance level) with slight modifications. This high sensitivity suggests a lack of robustness in ToM performance. GPT 4 outperformed its predecessor, GPT 3.5, showing improvement in ToM performance to some extent. The models generally performed better when tasks were presented in English than in Turkish. These findings indicate that GPT models cannot reliably pass first-order and second-order ToM tasks in either of the languages yet. The findings have significant implications for *Explainability* of LLMs by highlighting challenges and biases that they face when simulating human-like ToM understanding in different languages.

1. Introduction

As humans, we develop an understanding of other people's mental states (i.e., theory of mind), which enables us to attribute their behaviors to their desires, intentions, knowledge, and beliefs (Wellman and Liu, 2004). With the emergence of Large Language Models (LLMs) such as Open AI's ChatGPT, our everyday experiences are changing; we increasingly engage in prolonged dialogues with these virtual agents and assume that they understand our mental states. The question of whether these agents could exhibit theory of mind understanding—a process exclusive to humans—as well as the question of whether they can have artificial general intelligence (AGI), are currently topics of intense debate (Bostrom, 2014; Bubeck et al., 2023; Ullman, 2023; Kosinski, 2023). A recent method for

^{*} Corresponding author.

E-mail address: burcu.unluturk@yeditepe.edu.tr (B. Ünlütürk).

assessing LLMs' intelligence level is using existing psychological tests developed for humans (Bubeck, 2023; Binz and Schulz, 2023; Coda-Forno et al., 2023; Kosoy et al.; Van Duijn et al., 2023). These tests include vignettes, standardized measures and scales about human decision-making processes, verbal and mathematical reasoning, theory of mind (ToM), and emotional understanding. We believe that administering these tests to LLMs can provide us insight into their current capabilities and help us gain a deeper understanding of the complexities involved in human language and reasoning, especially in processing abstract situations and scenarios. This understanding is crucial for learning from others, making meaning, interpreting communication, and building rapport. However, most of this research is conducted in English, and limited research examines LLM's performance in different languages. To address this gap in the literature, in this study, we examine the performance of GPT 3.5 and 4 in first-order and second-order ToM scenarios prepared in English and Turkish. By simulating ToM judgments and explanations in GPT models and evaluating their explanatory quality and consistency, we aim to understand these models' behaviors, capabilities, and limitations.

2. Related work

How does ToM understanding develop in humans? Thus far, research in the field of developmental psychology shows that infants discern/perceive the associations between the goals/intentions of people and their actions (learning statistics) and can make predictions based on associations (Call and Tomasello, 2008; Ruffman, 2023; Woodward, 1998). It is possible to say that there are specific early capacities underlying ToM understanding, such as infants' sensitivity to human face configuration (e.g., Turati et al., 2006), interest in infant-directed speech (e.g., Cooper and Aslin, 1989), and engaging in joint attention (e.g., Akhtar and Gernsbacher, 2008; Carpenter et al., 1998). Building on these initial capacities, children actively construct naïve/informal theories of other people's mental states (desires, beliefs, knowledge, and emotions) as they explore the social world, and their understanding becomes more advanced over time with experience and learning (Gopnik, 2003; Gopnik and Wellman, 1994; Wellman, 2002; Wellman and Liu, 2004).

A very robust finding in the literature is that children pass first-order false belief tasks around 4-5 years of age. In these tasks, children are presented with a scenario where an agent has a false belief about an item's location or container's contents. Before four years of age, children disregard that the agent has a false belief and state their own knowledge of reality in these scenarios. However, after four years of age, they acknowledge that the agent has a false belief because they are unaware of the location change or the hidden contents (Wimmer and Perner, 1983; see Wellman, Cross and Watson for a meta-analysis). This developmental milestone is reported across different sociocultural contexts, suggesting a degree of universality. However, there is also research showing differences in the trajectory of ToM development due to cultural practices, differences in language use, and linguistic nuances (e.g., Bartsch and Wellman, 1995; Carpendale and Lewis, 2004; Naito and Koyama, 2006; Matsui et al., 2009; Peterson and Siegal, 1998; Vinden, 2001; Wellman et al., 2006).

Research in the Turkish sociocultural context demonstrates that Turkish preschoolers also pass first-order false belief tasks around 4-5 years of age, suggesting consistency in the performance of children acquiring the Turkish language with their Western English-speaking counterparts (Selçuk et al., 2018; Ünlütürk and Aksu-Koç, 2015). However, there are some differences observed in the developmental trajectory of ToM components, such as understanding diverse knowledge states earlier than diverse beliefs (Selçuk et al., 2018; Ünlütürk and Aksu-Koç, 2015), which is a pattern noted in collectivistic cultures before (e.g., Shahaian et al., 2014; Wellman et al., 2006). In addition, performance in false belief tasks is not only shaped by cultural practices but also associated mainly with specific linguistic features of a language. Some languages such as Turkish, Chinese, Russian, Puerto Rican Spanish, and Japanese use specific mental state verbs to mark the presence of false beliefs. Research shows that children acquiring these languages expect to hear these verbs in false belief scenarios. If questions in ToM tasks are not asked using the false belief verbs, they show lower performance in false belief tasks (Lee et al., 1999; Tardiff et al., 2004; Shatz et al., 2003; Matsui et al., 2009). Building upon these insights, the current study investigates how LLMs models trained in English vs. Turkish perform in false belief tasks. The distinct linguistic structure of Turkish with false belief verbs and the differences in how false beliefs are encoded in the training data culturally and linguistically provide a unique opportunity to examine the capabilities of LLMs in this domain.

It is also important to note that mastering first-order false belief tasks does not mark the end of the path toward comprehending other people's mental states. As children participate in the social world and grow older, their ToM understanding becomes more advanced and nuanced; they begin to reason about an individual's false belief about another individual's false belief around 6-7 years of age (second order ToM) (Wimmer and Perner, 1983; Perner and Wimmer, 1985; Perner and Howes, 1992). This developmental progression highlights the continuous advancement of ToM with maturation and experience and provides the developmental standards through which we can evaluate the capabilities of LLMs. In our exploration of first-order and second-order false belief understanding in LLMs, we aim to understand whether these models can navigate the intricacies of understanding other people's mental states similar to children. By placing the performance of LLMs in this developmental continuum, we seek to provide insights into the continuously changing capabilities of LLMs in their simulation of complex social reasoning tasks.

Considering that meaningful human interactions rely on ToM understanding, a way for LLMs to attain human-like reasoning and actively engage in meaningful social interactions with humans involves at least some level of proficiency in ToM understanding. As soon as advanced models of LLMs, like GPT, emerged, researchers began assessing the ToM capacities of its different models, and there is a growing debate about whether these LLMs show ToM understanding or not. For instance, two recent studies tested the performance of GPT-4 and its older versions (i.e., ChatGPT, GPT 3.5 turbo, text-davinci-003), using scenarios based on classic false belief tasks and concluded that GPT-4, by far, showed the best performance in giving the most insightful and resourceful responses compatible with these scenarios and thus show development of ToM understanding (Bubeck et al., 2023; Kosinski, 2023). GPT-4 has been exposed to extensive datasets containing information on mental states (e.g., stories, fables, narratives). This exposure may have equipped the

model with the capacity to comprehend and produce recursive sentences, such as “Josh thinks that Anne believes that Jack knows that she is in Paris.” Kosinski asserts that this successful grasp of ToM scenarios is a natural outcome of the impressive level of human-like language skills in GPT-4.

From a different perspective, Ullman (2023) shows skepticism about the robustness of Kosinski’s findings and focuses on when and how the GPT models fail in the ToM tests. For this end, he slightly modifies the novel Unexpected Contents and Unexpected Location scenarios Kosinski used and tests them with GPT 3.5. Ullman made three modifications to the Unexpected Contents scenario: 1) *transparent access* (the bag is transparent so anyone can see what is inside), 2) *uninformative label* (the character cannot read the label), 3) *trusted testimony* (a trusted friend of the character tells the content of the bag), and 4) *late labels* (the label is put afterward and indicating the wrong content). Ullman also made three modifications to the Unexpected Transfer scenario: 1) *transparent access* (the containers are see-through so anyone can see the cat), 2) *relationship change* (instead of “in,” “on” prepositions are used to indicate the location of the cat), 3) *trusted communication* (the second character calls and informs the first character about the location transfer), and 4) *the mental state of the other character* (asking what the second character thinks about the location). These modifications significantly decreased the performance of GPT 3.5 in these ToM scenarios. Based on these findings, Ullman argues that even slight deviations from the common scripts can dramatically deteriorate the performance of these models, thus, their performance in ToM tasks is not robust yet. These findings highlight the need for improving the flexibility of the GPT-3.5 for different ToM scenarios and call for careful examination of the current limitations of LLMs in showing ToM understanding.

Overall, studies by Kosinski and Ullman reveal a notable discrepancy regarding the ToM capabilities of different GPT models 3.5 and 4. Such inconsistencies not only point to the improvement in the ToM performance in new GPT models but also raise important questions about the impact of linguistic context and content/nature of the given task/prompt on the ToM performance of LLMs. Moreover, acknowledging that these models are predominantly trained on vast amounts of English data, and yet their proficiency is rapidly increasing beyond the realm of English (e.g., Penedo et al., 2023), in the current study, we aim to investigate the theory of mind (ToM) performance of GPT 3.5 and 4 in Turkish and English scenarios. This comparative analysis in two languages enables us to examine the ToM capabilities of these models beyond the realm of English language proficiency.

In addition, recent research provides an alternative perspective and considers these agents as cultural products/technologies that represent the knowledge accumulated in a given language. As a result, these technologies can serve as tools for imitation and linguistic and cultural transmission (Kosoy et al., 2023; Yui, Kosoy and Gopnik, 2023). For instance, Kosoy et al. examined Google’s LLM LaMDA’s ToM capabilities using social cognition and ToM tasks from the developmental psychology literature as well as including Ullman’s variations of these tasks. They found that LaMDA probability of responding correctly in these tasks were at best at chance level. Moreover, LaMDA does not necessarily follow the human developmental trajectory in these tasks by showing high performance in verbal tasks but low performance in tasks that require real-world interaction and exploration.

Building on these studies, we believe that comparing the ToM performance of LLMs in different languages can shed light on whether training data and language structures can lead to nuances in the performances. This comparison is particularly intriguing because LLMs can learn the statistical patterns in large data (just like infants and young children learn associations between intentions and behavior). However, unlike infants and young children, these models may only replicate statistical patterns rather than showing actual comprehension. Examining GPT models’ performance in English and Turkish can reveal how training data and language structure may influence their ability to understand and perform in complex ToM scenarios.

Thus, this study had three overarching goals:

- (1) The first goal is to replicate previous findings reported by Ullman (2023) by assessing the ToM performance of not only GPT 3.5 but also GPT 4 in English and Turkish (Study 1) using first-order false belief tasks and thereby testing the generalizability of previous findings with a more sophisticated GPT model in a different linguistic context. Ullman’s (2023) study used GPT 3.5, and it is currently unknown whether a more advanced GPT Model (GPT-4) can pass the variations used by Ullman as GPT-4 has more parameters and, as a consequence, has better knowledge representation and complex reasoning skills (Open AI Research, 2023). It is also currently unknown whether and to what extent GPT-3.5 and GPT-4 models show ToM performance in a language other than English. Turkish is a difficult-to-model language with a different syntactic and morphological structure compared to English, and it has a smaller digital footprint than English; thus, have smaller training data.
- (2) The second goal is to extend this replication by eliciting responses from GPT-3.5 and GPT-4 models via first-order false belief scenarios within the same chat session that includes both close-ended prompts and open-ended explanation questions across two languages (Study 2). This approach goes beyond the existing literature by evaluating/coding the GPT-3.5 and 4’s responses to open-ended explanation questions and provides us insights into the depths and quality of ToM reasoning these models can provide within a prolonged conversation.
- (3) The third goal is to examine the performance of GPT-3.5 and GPT-4 models in more complex second-order false belief scenarios across two languages (Study 3) (Wimmer and Perner, 1985; Arslan et al., 2017; Hollebrandse et al., 2011). This way, we examine the models’ capacity to track more than one false belief in intricate scenarios via a variety of close-ended prompts and open-ended questions during a prolonged conversation within the same chat session.

In these studies, we expect that GPT-4 may outperform GPT 3.5 first-order and second-order ToM tasks due to its training via feedback from humans, a higher number of parameters, and improved human-like reasoning capabilities with real-world use compared to its predecessor, GPT-3.5, (Open AI, 2024). Our expectation is based on a developmental psychology framework where higher verbal skills, familiarity with mental state talk, and experience in the social world were associated with better ToM performance in children (e.g., Wellman et al., 2001; Dunn and Brophy, 2005). An important contribution of this study is examining the GPT model’s responses

to open-ended explanation questions. This examination will not only allow us to examine the accuracy of final judgments but also the reasoning processes of the models. The effect of language features on ToM performance will be explored. We included the two versions of the false belief verb *san-* and *zannet-* as they are semantically similar and interchangeably used in everyday language (retrieved from tdk.gov.tr). As mentioned earlier, previous research indicates a link between the absence of these verbs and diminished ToM performance (Lee et al., 1999; Shatz et al., 2003; Matsui et al., 2009). Given these prior findings, we investigate whether this potential link holds true in our study with Turkish, which has these false belief verbs.

It is also important to note that, in these studies, comparing English and Turkish provides a unique context for several reasons. First, English presents an advantage over Turkish with regard to language statistics due to the volume and the variety of data available on the web for model training. As the dominant language of digital platforms, English has the benefit of large datasets, including language samples with different styles, topics, and contexts. This enables a more comprehensive statistical representation of the language that supports human-like language understanding and production. On the other hand, Turkish models are trained on relatively smaller datasets showing less diversity. Thus, LLMs may be as proficient and fine-grained when operating in Turkish as when operating in English. This could influence the models' performance in natural language processing and cognitive tasks such as ToM.

Second, Turkish and English come from different language families and have distinct syntax and morphology. English strings the words as Subject-Verb-Object, sentences are completed with relative clauses and mental state verbs such as "think" or "believe" can be used in such complement clauses. On the other hand, Turkish follows the Subject-Object-Verb order, but it is free to vary due to context-based agglutinating structure. Turkish has a complex verb system with tense, aspect, and modality inflections (Göksel, 2004). It is a pro-drop language meaning that you do not need to use the subject of the sentence all of the time. You can drop the subject pronouns as the verb is inflected according to the subject of the sentence. Moreover, Turkish is also characterized by the absence of auxiliary verbs (e.g., am, is are) and definite/indefinite articles (e.g., a, the). The definite/indefinite distinction of nouns are mainly inferred from the context (Ketrez, 2012). In Turkish, it is possible to produce very long words by joining many suffixes to the word. For instance, "evlerinizden" is a single word in Turkish but it corresponds to three words in English "from your houses" with *ev* meaning house, *-ler* is suffix for plural, *-iniz*, suffix for possessive and *-den* is suffix for dative (Ketrez, 2012). Turkish has vowel harmony meaning that words follow a pattern when they are strung together, and the vowel of suffixes can also change depending on the vowels in a word. Thus, it is harder to determine the word stems in Turkish than in English. All these characteristics of Turkish make it a difficult model language for AI-based conversational agents because it is harder for LLMs to learn and predict the distributional/statistical properties of a language like Turkish than English (Ofaşer and Saraçlar, 2018). This difficulty can influence the performance of the model in Turkish during specific language-related and cognitive tasks.

Third, in addition to its syntactic and morphological differences, Turkish has specific false belief verbs *san-* and *zannet-* denoting that a person has a false belief, which may impact ToM performance. As mentioned before, studies show that explicit marking of false beliefs via verbs in some languages enhances children's performance in ToM tasks. Thus, we explore whether a similar pattern is observed in LLMs when ToM prompts/questions are given using false belief verbs in Turkish. We used both false belief verbs *san-* and *zannet-* in Turkish. Although their meaning is the same within the context of false belief scenarios, *zannet-* is an older verb with Arabic origin, while *san-* is a Turkish verb that has several meanings like assume, think, and false believe, depending on how it is used in different contexts. Thus, by exploring LLMs' ToM understanding in Turkish, we aim to provide insights into whether language structure influences information representation and reasoning processes in these models.

3. General methodology

3.1. Simulation setup and parameters

For this study, we ran simulations using Open AI's Chat Completion API (Open AI, n.d.) via Python software (version 3.11.2) via Jupyter Notebook. We used GPT-3.5-turbo-0613 and GPT-4-1106-preview models. In all simulations, we set the system message to "You are a helpful assistant," as in the previous studies (Binz and Schulz, 2023; Coda-Forno et al., 2023; Kosinski, 2023; Ullman, 2023). This approach not only enabled us to receive consistent responses from the GPT models across all scenarios but also facilitated direct comparison with previous studies and replicability of Kosinski's and Ullman's findings. Moreover, we arranged the temperature parameter to 0 to receive consistent responses and for replicability. Temperature parameter (ranging from 0-2) controls the randomness of responses to be elicited from these models, with lower temperatures yielding more consistent and predictable responses than higher temperatures. Thus, by setting the temperature to 0, we prioritized the replicability of our study's results by other researchers. We also aligned our methodology with Kosinski's and Ullman's previous studies to have comparable results contributing to understanding of LLMs' capabilities. All other parameters remained as default.

3.2. Experimental design and predictors

We used GPT models to simulate responses from multiple human participants and administered ToM tasks with children. We conducted at least 10 trials for each task as if we had 10 participants. The idea is to evaluate how human-like responses GPT models can give and whether the errors they make resemble to humans. This is a new methodology (similar to earlier Turing Tests) and is increasingly gaining attention in the field (e.g., Aher et al., 2023; Binz and Schulz, 2023; Coda-Forno et al., 2023; Jiang et al., 2022). Our factors/predictors in our experimental design included task language {0: 'English,' 1: 'Turkish'}, the ToM task category {0: 'Unexpected Contents,' or 1: 'Unexpected Transfer'}, GPT model {0: 'GPT-3.5-turbo-0613', 1: 'GPT-4-1106-preview'}, language verb category {0: 'English,' 1: 'Turkish,' 2: 'Turkish san-', 3: 'Turkish zannet-'}, and ToM question type: {0: 'Action,' 1: 'Belief,' 2:

'Explanation,' 3: 'Knowledge,' 4: 'Reality', 5: 'Thought'). Thought and Belief ToM question types are first-order false belief questions that are either asked with the verb "believe" or "think." In the Figures in the Results section displayed them as first-order false belief questions for ease of presentation. Python codes and data structure parameters for running simulations are available at this link: https://github.com/onurbal101/gpt_tom.

3.3. Data collection and analysis

We used two methods for collecting data from GPT models: 1) text sampling analysis and 2) log probability analysis. When GPT-4 was first made available, and at the time of our data simulation, there were limitations in accessing log probability endpoints in Open AI's Chat Completion API (Open AI Community, 2023). Thus, we decided to use text sampling and coded the linguistic data/responses based on accuracy and response quality manually. This approach yielded rich qualitative data that reflected the narrative coherence and complexity of the responses given by the models. Second, we used log probability analysis as soon as it was made available again to supplement the text sampling method. We describe the details of each method below.

3.3.1. Text sampling

Using this method, we ran 10 trials of each ToM task, including different prompts/questions. Then we recorded the generated text and coded it for response accuracy and explanation quality. Each response was coded using a binary scheme where "0" indicated an incorrect response or inadequate explanation, and "1" indicated a correct or adequate explanation. As our outcome was binary, we conducted mixed effects logistic regression models using R. This analysis allowed us to estimate the effects of multiple factors on outcomes and understand the odd ratios of correct responses.

The general formula for the models is as follows: $\text{logit}(P(\text{is_correct} = 1 | b_i, X_i)) = \beta_0 + \beta_1 \text{story_category}_i + \beta_2 \text{story_language}_i + \beta_3 \text{model}_i + \beta_4 \text{question_tom_type}_i + b_{0i}$

In this formula, $\text{logit}(P(\text{is_correct} = 1 | b_i, X_i))$ represents the log-odds of the correct response ($\text{is_correct} = 1$). $\beta_0, \beta_1, \beta_2, \beta_3, \beta_4$ are the fixed effects coefficients, representing the effect of each predictor variable (task_category, language, model, question_tom_type) on the log-odds of answering correctly. Then, b_{0i} represents the random effect coefficient. When using glmer function in R, we first took chat session ID as random effects since we asked all the questions within the same chat session (nested data in repeated design). However, the boundary fit was singular suggesting that there is almost no variation across sessions and random structure may not be necessary in the model. Thus, we built a simpler model by excluding random effects. Finally, task_category_i, language_i, model_i, question_tom_type_i represent the predictor variables for observation i.

We used a subset of these predictors in each study we reported here. We describe the details of each study below. For verb category, we conducted chi square analysis within Turkish data to compare the performance across different verbs in Turkish. The data for each study and the R codes we used for analysis and drawing graphs are available at Open Science Framework at this link: <https://osf.io/jysxt/>

3.3.2. Log probability analysis¹

Once the log probability endpoints become available in Chat Completion Endpoints, we integrated this methodology to complement our text sampling analysis. We used this analysis only for the close-ended prompts/questions. Log probabilities or "logprobs" enabled us to quantitatively evaluate the likelihood of correct responses. During text sampling, while we perceive the response being received as a whole (such as "chocolate"), the ChatGPT language model is actually responding in chunks, called tokens (such as "choc" and "olate"). Each token's likelihood of being picked by the language model is returned by the API in the form of its logarithmic equivalent, called its logarithmic probability, or logprob for short. This value is between negative infinity and 0. To calculate a more easily interpretable probability score between 0% and 100%, taking the exponential of this value suffices. While knowing the respective probabilities of the tokens "choc" and "olate" may be useful, we were primarily interested in the likelihood of the entire response, "chocolate", being received. To receive a complete standardized response, we updated our systems message as "You are a helpful assistant. Respond with the digit corresponding to the correct answer to replace {RESPONSE}," we also updated our prompts to have forced-choice digit response options as (1) or (2). This way the model responded by either selecting option 1 or 2. The binary choice approach is the most commonly used methodology in the literature with these types of cognitive tasks due to its simplicity and ease of evaluation (e.g., Binz and Schulz, 2023, Kosinski, 2023). We also preferred using this approach when evaluating the models' capability of making correct choices because the models were more likely to provide consistent and reliable responses that can be directly comparable across trials. Considering the variability of the log probabilities in each trial, we ran 10 trials and averaged the log

¹ We conducted an additional analysis using the surprisal-based method, which involves summing the log probabilities (log probs) of the token sequence for multi-token critical words to complement our binary choice method and test the robustness of our findings. We calculated the cumulative surprisal score and compared the log probability of correct responses by language, model, and task. This analysis yielded higher variability in responses which makes interpretation of the findings difficult. However, the overall pattern of results was similar to our findings with the log probability analysis using the binary choice option, showing performance at chance levels. Although there were some variations across tasks, GPT-4 performed better than GPT-3.5. Unexpected Contents yielded correct responses with higher confidence than the Unexpected Location Transfer Task, and performance in English was better than in Turkish. Moreover, in Study 2, where the questions were presented within the same chat session, the models had higher confidence than in Study 1. In Study 3, the models showed high confidence in second-order ToM questions when compared to the variations of ToM tasks in Study 1 and Study 2. The data and the summary of the results can be found in the OSF link: <https://osf.io/jysxt>.

probabilities of correct responses to reliably assess the models' performances. We provided Log Probability Results as Figures within the text, and we included exact probabilities in Tables in Appendix B.

In each study reported below, we first provide results from our text analysis including both close-ended and open-ended prompts, and next, we provide results from log probability analysis only for close-ended prompts.

4. Study 1: Ullman (2023) replication in English and Turkish

In Study 1, we replicated Ullman's (2023) study using identical task versions and chat completion prompts given as part of a separate chat. We extended the replication by including Turkish, using two GPT models (GPT-3.5 and GPT-4), and varying the prompts using false belief verbs *san-* and *zannet-* to explore linguistic influences. Based on Ullman's findings, we hypothesize that both GPT models would fail (perform below chance levels) in variations of basic ToM tasks, while GPT-4 would perform better than GPT-3.5 (Kosinski, 2023; Bubeck et al., 2023). The models would perform better in English compared to Turkish due to the more extensive and varied training data available in English. We hypothesized that GPT models would respond to false belief questions more accurately when questions were phrased using false belief verbs instead of general cognitive verbs. We did not expect any significant difference between the Turkish false belief verbs *san-* and *zannet-* as they are semantically similar and can be used interchangeably.

4.1. Materials and prompts

4.1.1. Unexpected contents task

We used the baseline story and its versions as 1A) *Transparent Container*, 1B) *Uninformative Label*, 1C) *Trustworthy Testimony*, and 1D) *Late Labels* (Ullman, 2023). In the baseline version, a character called Sam finds a popcorn-filled bag. The bag does not include any chocolate, but its label says "Chocolate" by mistake. Sam cannot see the contents of the bag but reads the label. In the baseline version, we ask GPT what Sam would think about the contents of the bag when she sees it. In 1A) *Transparent Container* version, GPT is informed that the bag is transparent and asked what Sam would think about the bag's contents when she sees the bag. In 1B) *Uninformative Label* version, Sam cannot read the label, so the label does not have any informative value. In 1C) *Trustworthy Testimony*, Sam's friend calls and informs Sam about the contents of the bag despite the misleading label, and Sam trusts her friend. We again ask GPT what Sam would think about the contents of the bag when she sees it. In 1D) *Late Labels*, Sam herself fills the bag with popcorn and then puts a label on it that misleadingly says chocolate. In this case, we asked GPT what Sam would think about the contents of the bag. For Ullman's (2023) replication, two prompts were given to GPT in separate chat sessions:

- (1) She believes the bag is full of {RESPONSE} (belief prompt)
- (2) She is delighted to have found this bag. She loves eating {RESPONSE} (action prompt)

4.1.2. Unexpected location transfer task

We used the baseline ToM task and its versions as 2A) *Transparent Access*, 2B) *Relationship Change*, 2C) *Trustworthy Testimony*, 2D) *Additional Person* (Ullman, 2023). In the baseline task, there is a cat in a room with a box, a basket, and two characters, John and Mark. John places the cat in the basket, and while he is away, Mark places it in the box and leaves, at which point John returns to the room. We asked GPT where John would look for the cat in the room. In version 2A) *Transparent Access*, both containers are see-through (a glass chest and a plastic box), so the cat should be visible to John regardless of its location. In version 2B) *Relationship Change*, the cat is not placed inside an object but is placed on top of an object and thus should be visible to the characters. In version 2C) *Trustworthy Testimony*, Mark calls John and tells John about the location change, and John trusts Mark. In 2D) *Additional Person*, the baseline version of the story is given, but instead of John, the mental state of Mark is inquired, who is not supposed to be holding a false belief. For Ullman's (2023) replication, two prompts were given to GPT in separate chat sessions for the baseline and the first three variations:

- (1) John thinks that the cat is in the {RESPONSE} (belief prompt).
- (2) John will look for the cat in the {RESPONSE} (action prompt)

As for the fourth variation (2D), the prompts were:

- (1) Mark thinks that the cat is in the {RESPONSE} (belief prompt).
- (2) Mark will look for the cat in the {RESPONSE} (action prompt)

We adapted these scenarios into Turkish and varied the verb type as "believe", "*san-*" and "*zannet-*" in Unexpected Contents and "think", "*san-*" and "*zannet-*" in the Unexpected Location Transfer. (See Appendix A in Supplementary Materials for all scenarios and prompts in English and Turkish).

4.2. Data preparation and coding

For this study, we repeated each prompt 10 times across GPT models in English and Turkish, which yielded a total of 1200 responses. We generally received one-word utterances when we used chat completion prompts with {RESPONSE}. To ease the data coding process, we had a column for correct responses (i.e., answer key), and we saved the elicited responses on another column next to

the answer key on an Excel sheet. We used the Python equality operator to compare the correct answer in the key column with the elicited response. If the responses were identical, the result yielded “true,” if not, “false.” For responses with more than one word, we first determined unique responses and then coded them manually. Since the temperature was set to 0, there was variety, but there were also some repetitive responses. Once we coded a unique response, we matched it with the other identical responses with more than one word or phrase. For manual codes, both authors coded all responses for inter-reliability. The percentage of agreement for all correct responses was 98%, and Cohen’s kappa = .97 shows very high agreement.

4.3. Results

4.3.1. Text sampling results

GPT models were 40% (N=476) correct in responding to first-order ToM prompts. While both GPT models were 100% correct in base stories showing ceiling performance, the accuracy of responses dramatically dropped in the different versions and showed a pattern similar to Ullman’s (2023) findings (see Fig. 1 for the percentage of correct responses in each task for each prompt across trials). We conducted logistic regression analyses using RStudio on the accuracy of the responses provided by GPT (true = 1; false = 0). For the analysis, we used base R functions and the plyr package (Wickham, 2011) for data preparation and the lme4 package (Bates et al., 2015), and the car package (Fox and Weisberg, 2019) for the analysis. In our model, we include Language (English vs. Turkish), Story Category (Unexpected Contents vs. Unexpected Transfer), Model (GPT-3.5 vs. GPT-4), and ToM Question Type (Action, Belief, and Thought) as predictors. The findings showed that GPT models were less accurate in responding to Unexpected Transfer Task prompts than Unexpected Contents Task prompts ($OR = .55$, $\chi^2(1, N = 1200) = 8.02$, $p < .001$). There was a significant effect of language on GPT models’ likelihood of giving responses ($OR = 1.69$, $\chi^2(1, N = 1200) = 15.41$, $p < .001$). Contrary to our expectations, GPT provided more correct responses to ToM prompts in English than in Turkish. Model also showed significant effect ($OR = 1.65$, $\chi^2(1, N = 1200) = 17.5$, $p < .001$). GPT-4 was more likely to give accurate responses compared to GPT-3.5. ToM Question type (action, belief, and thought) did not have any effect on the accuracy of responses (see Fig. 1).

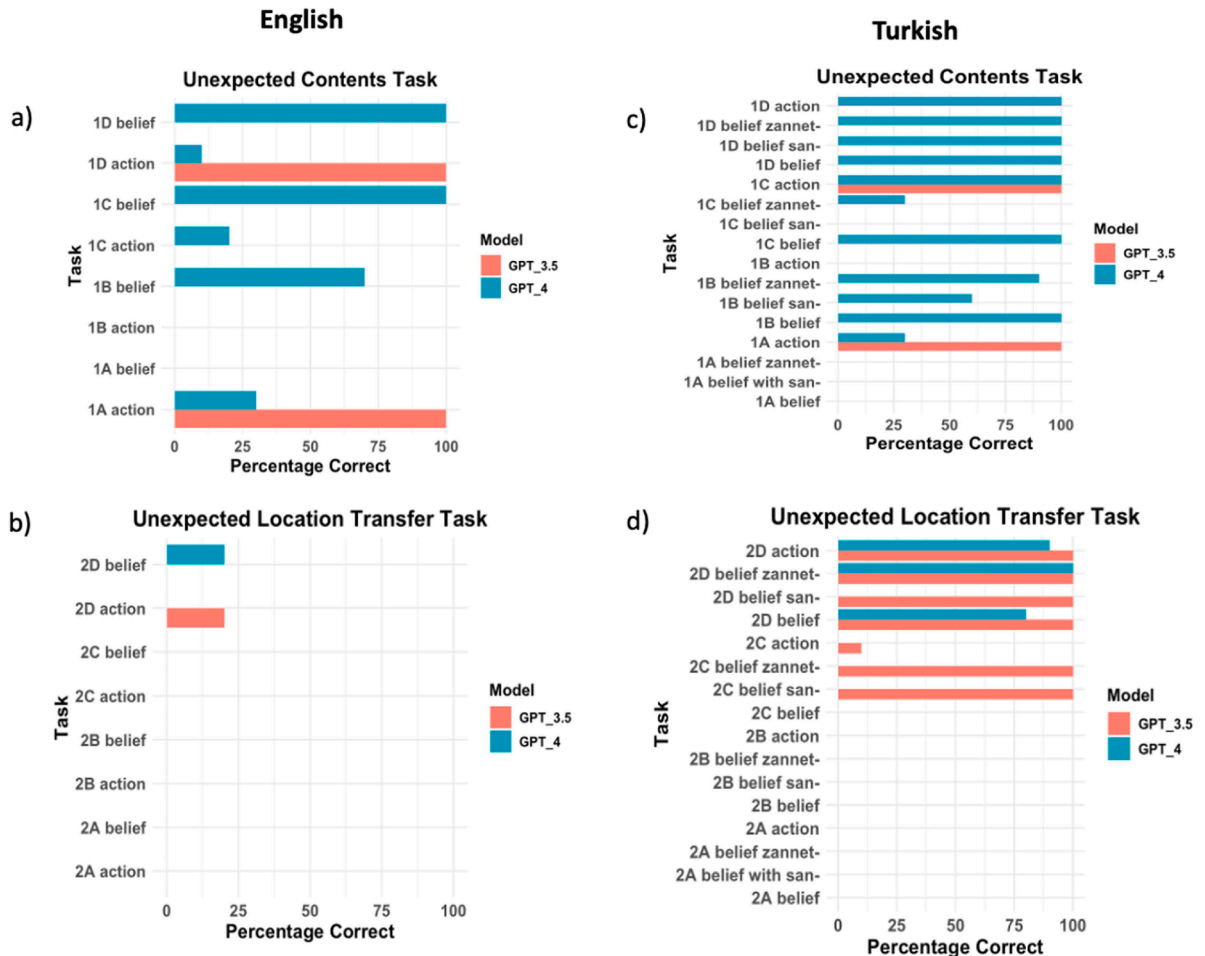


Fig. 1. Accuracy of responses in first-order ToM tasks across language and GPT models with text sampling in Study 1.

Next, we zeroed in on belief prompts to examine whether there were any differences between English and Turkish false belief verbs. We conducted a chi-square analysis on the frequencies to determine whether false belief verb type: 0 = English (believe or think), 1 = Turkish (believe or think), 2 = Turkish *san-*, 3 = Turkish *zannet-* is associated with the accuracy of responses. We found a significant association, $\chi^2(3) = 19.27, p < 0.001$. We conducted pairwise comparisons using Holm correction to examine which verb types yielded more accurate responses. We found that Turkish general cognitive verbs yielded more correct responses than English cognitive verbs, $p < .001$. We also found that the Turkish false belief verb *zannet-* yielded more accurate responses than English cognitive verbs, $p < .01$ (see Fig. 2).

As part of our qualitative analysis of the data, we realized that while we included only close-ended prompts when inquiring about ToM understanding, GPT models, particularly GPT-4, provided us with some speculative responses and went beyond the scope of our question, and the information given in the scenario when responding to the prompts. For instance, in scenario 1B, where Sam is unable to read the label, so the label does not have any informative value, GPT-4 responded to the prompt “She believes that the bag is full of {RESPONSE}” by speculating about clues that could lead to understanding what is inside the bag such as the appearance/shape of the bag, pictures on it, or its smell. The frequencies of occurrences for going beyond scope were 70% for the belief question in 1B, 50% for the action question in 1C, 40% for the action question in 1D, and 30% for the thought question in the 2D scenario in GPT-4’s responses in English. Going beyond scope was less prevalent in Turkish, but we have observed 30% of GPT-4’s responses going beyond scope in scenario 1B. This finding highlights the inferential/speculative thinking capabilities of the more advanced GPT-4 model, but still, the performance was not consistent enough to enable this model to pass first-order ToM tasks.

4.3.2. Log probability results

The log probabilities of correct responses showed that in the baseline belief and action questions, GPT-4 showed higher confidence (log probability 1.00 or close) than GPT-3.5 in its responses (see Fig. 3). On the other hand, GPT-3.5 showed moderate confidence even in the baseline prompts. Overall, models showed higher confidence in English, but the results were mixed in Turkish. In the variations of the tasks, while GPT-3.5 continued to show a moderate level of confidence, GPT-4 showed inconsistent performance. Compared to belief prompts, performance in action prompts was lower, with log probabilities dropping to 0.00 in some cases. This inconsistency showed that the task variations influenced GPT-4’s confidence levels dramatically, which could be due to the model’s sensitivity to contextual and linguistic differences. We also observed that false belief verbs *san-* and *zannet-* in belief prompts did not systematically influence models’ performance. Moreover, similar to text sampling results, GPT-3.5 and GPT-4 showed higher performance in the Unexpected Contents Task than the Unexpected Location Transfer Task. These findings highlight the complex capabilities and limitations of LLMs in processing ToM prompts due to task demands and language-specific cues.

Through both text sampling and log probability results, we replicated Ullman’s findings. GPT-3.5 and GPT-4 perform only at chance level in the slight variations of first-order ToM tasks. Both text sampling and log probability results showed that GPT-4 outperformed GPT-3.5 across all prompts. Both models performed better in Unexpected Contents Tasks than in Unexpected Location Transfer Tasks. Language impacts models’ performance but the specific patterns differ across text sampling and log probability methods, thus we cannot talk about a consistent and robust pattern. Log probability analysis showed that GPT-4 showed higher

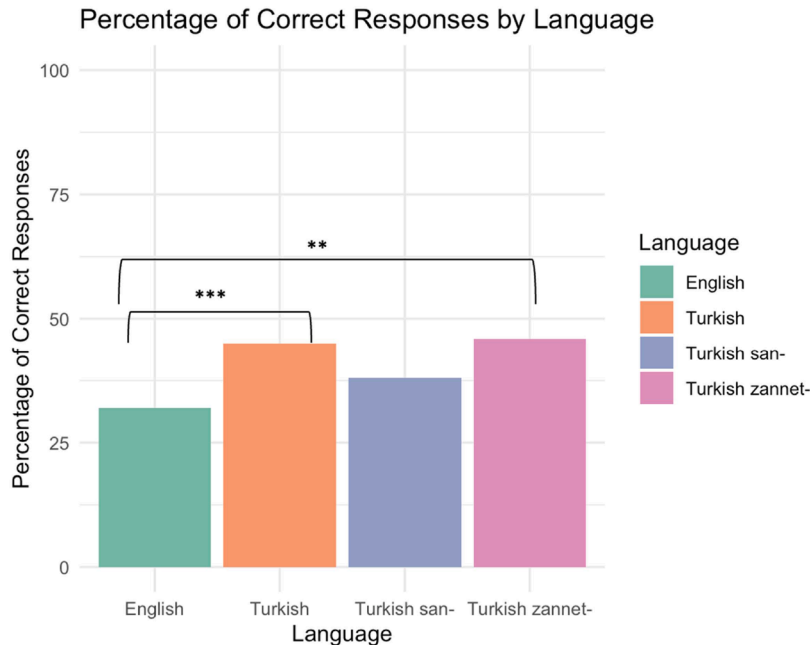


Fig. 2. Percentage of correct responses across different verb types in Study 1.

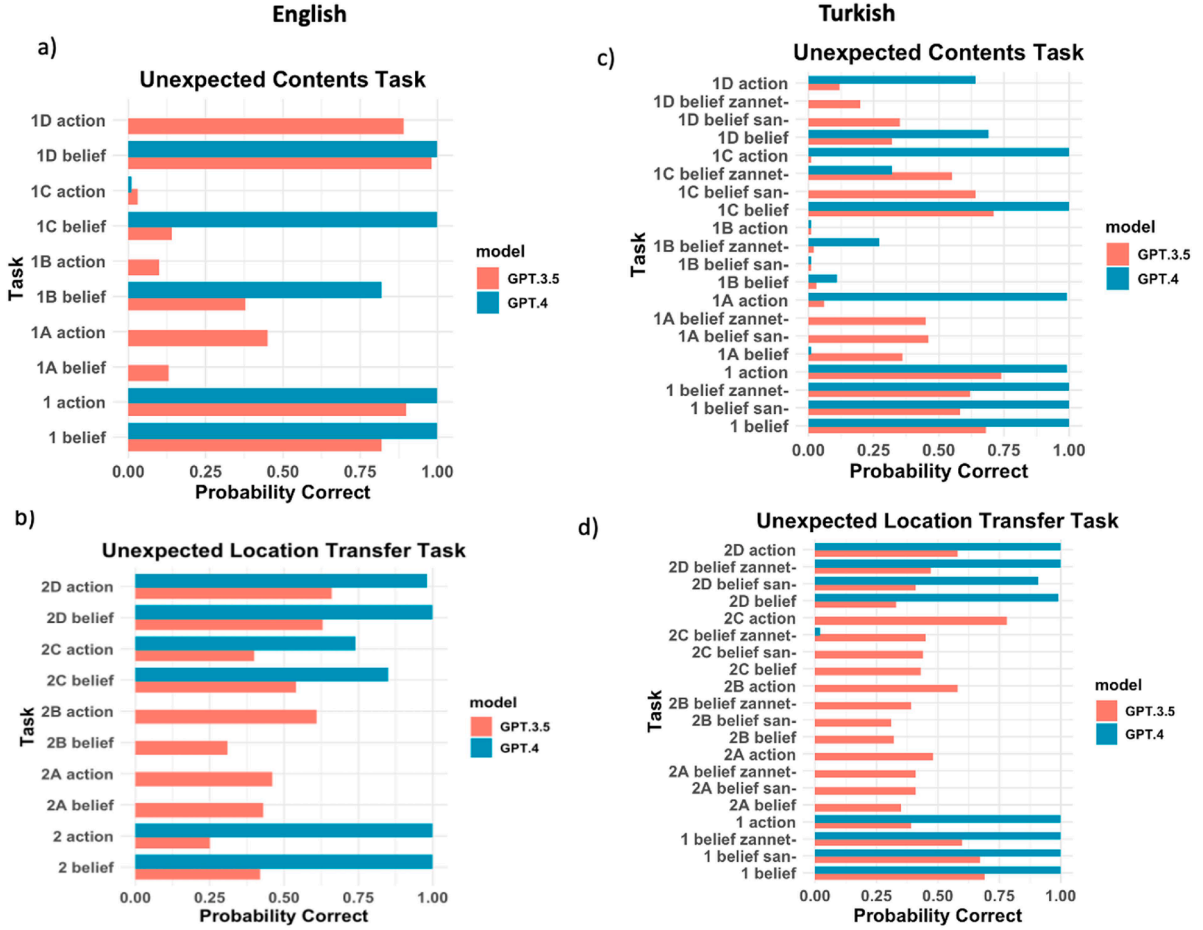


Fig. 3. Probability of correctness in the ToM tasks across language and GPT models in Study 1.

confidence in its responses; however, prompt content affected the response accuracy. For instance, some action prompts yielded 0 in confidence. Text sampling results highlights GPT-4's capacity to provide speculative responses to "why" questions. While text sampling results showed that "zannet-" verb in Turkish yielded more correct responses than in English, this pattern was not existent in log probability analyses.

5. Study 2: Ullman (2023) expansion in English and Turkish

Although by replicating Ullman's (2023) findings, we gained insight into the performance of different GPT models via the fill-in-the-blank type of prompts and receiving only one-word utterances, these findings are still limited with regard to metalinguistic judgments (Hu and Levy, 2023) and the inferential reasoning processes during a prolonged conversation within the same chat session. For this purpose, in Study 2, we provided prompts in the question form to GPT and asked a reality check/knowledge question adapted from Kosinski (2023). We added an explanation question to understand how GPT models provide reasoning for their decision, whether its reasoning is correct, and whether it goes beyond the scope of the information given when answering an open-ended explanation question. Moreover, we have given all prompts and questions within the same chat session to examine the effects of dynamic/interactive dialogue similar to real-life interactions with GPT models.

5.1. Materials and method

5.1.1. Unexpected contents task

We used the same Unexpected Contents Task with its versions of 1A) *Transparent Container*, 1B) *Uninformative Label*, 1C) *Trustworthy Testimony*, and 1D) *Late Labels*. This time, we had four prompts/questions in total, given within the same chat session.

- (1) She believes the bag is full of {RESPONSE} (belief prompt).
- (2) She is delighted to have found this bag. She loves eating {RESPONSE} (action prompt).

- (3) Why does she think that the bag is full of that item? (explanation question).
- (4) She opens the bag and looks inside. She can clearly see that it is full of {RESPONSE} (reality check prompt).

5.1.2. Unexpected location transfer task

We also used the same Unexpected Location Transfer Task with versions of 2A) *Transparent Access*, 2B) *Relationship Change*, 2C) *Trustworthy Testimony*, 2D) *Additional Person*. This time, we had 4 prompts/questions in total and they were given within the same chat session.

- (1) John/Mark thinks that the cat is in the {RESPONSE} (belief prompt).
- (2) John/Mark will look for the cat in the {RESPONSE} (action prompt).
- (3) Why will he look for it there? (explanation question).
- (4) The cat jumps out of the {RESPONSE} (reality check prompt).

5.2. Data preparation and coding

We used the same coding scheme as Study 1. In the next column, we compared the answer key column with responses elicited from GPT models. For long responses including more than one word or phrase, we again determined unique responses, coded them manually, and matched them with other identical responses. For open-ended explanation questions, we used a coding scheme examining 1) accuracy, referring to whether GPT reached the correct conclusion/response, 2) correct reasoning, referring to whether the thinking process that GPT uses is logical and coherent; and 3) going beyond scope for answers that included inferences or background knowledge to provide a plausible explanation for the question. For instance, in 1B) *Uninformative Label* scenario, the correct answer for “Why does she think that the bag is full of that item?” is that “Because Sam cannot read and cannot see what is inside

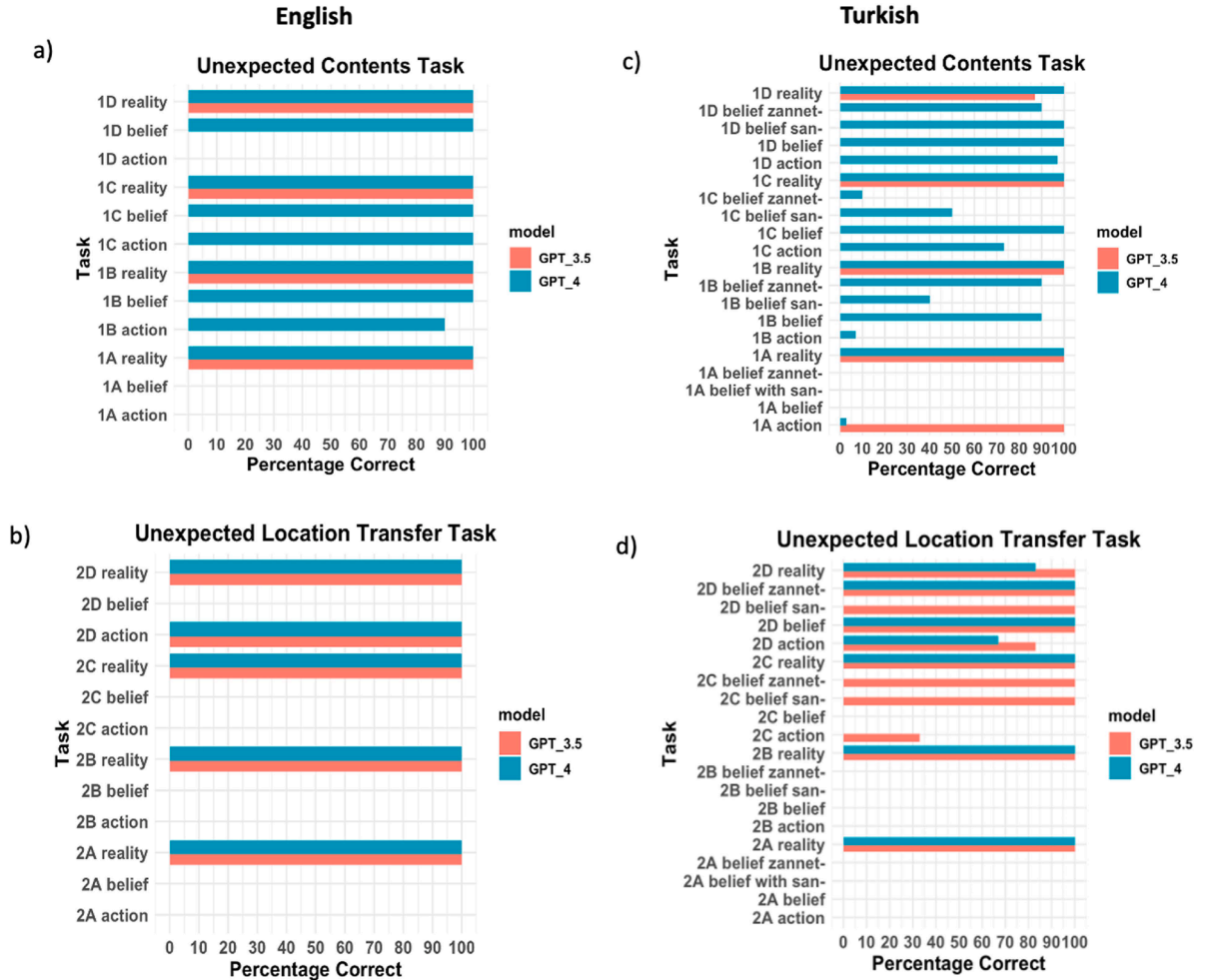


Fig. 4. Accuracy of responses in first-order ToM tasks across language and GPT models with text sampling in Study 2.

of the bag, we don't have a concrete idea as to what Sam might be thinking.” If GPT gives a response similar to the correct answer, it is coded as accurate. If the premises were correct in the explanation, such as saying that if Sam cannot read and if there is no other context provided to determine the contents of the bag, Sam cannot know what is inside, then the explanation is coded as having correct reasoning. Finally, if GPT offers more information to figure out what is inside the bag, such as checking its weight, its scent/smell, sound, weight and feel, or other visual and contextual cues, then we coded it as going beyond the scope of the question. In such cases, GPT provided a lot of additional information. For manually coded explanation-type open-ended responses, we calculated interrater reliability and found 95% agreement for having correct reasoning and 97% agreement for going beyond scope. Cohen's kappa values were .91 and .84, respectively, indicating high agreement between coders.

5.3. Results

5.3.1. Text sampling results

For Study 2, we elicited a total of 3200 responses. Of all these responses, 1756 (55%) were correct. Comparing the accuracy of responses in Study 1 (Fig. 1) and in Study 2 (Fig. 4) indicates that asking ToM questions within the same chat session slightly improved the GPT models' performance. Particularly, GPT-4 had ceiling performance in some versions of the tasks, while it showed much lower performance when questions were asked in separate chats. This observation shows that the interactive nature of the dialogue and building upon context, even in a short chat session, is important for GPT models to have better performance in ToM.

Next, we built logistic regression models to examine the effects of our predictors on the accuracy of responses. We used the same R packages as in Study 1. We included Language (English vs. Turkish), Story Category (Unexpected Contents vs. Unexpected Transfer), Model (GPT3.5 vs. GPT4), and ToM Question Type (Action, Belief, Thought, Reality) as predictors. We included session ID as random effects since we asked four questions in a row within a chat session. The model did not converge in the first run as our dataset is quite large and complex. Thus, we included a “bobyga” optimizer to run more iterations. The final model showed that, similar to Study 1, GPT models were less accurate in responding to Unexpected Transfer Task prompts than Unexpected Contents Task prompts ($OR = .39$, $\chi^2(1, N = 3200) = 9.85$, $p < .001$). There were not any significant differences across languages (Turkish vs. English), $p = .21$. GPT-4

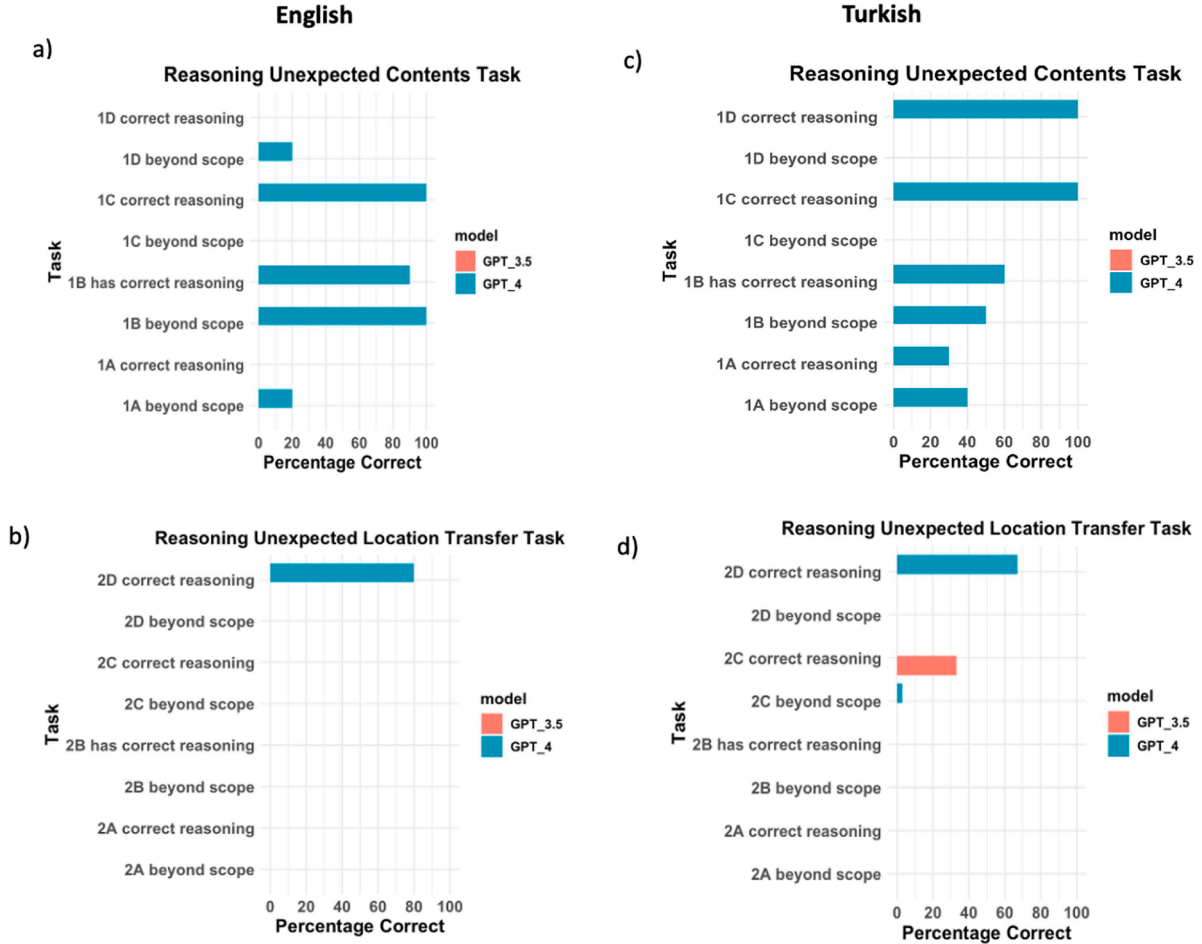


Fig. 5. Quality of responses to first-order explanation questions in Study 2.

was more likely to give accurate responses to ToM questions than GPT 3.5 ($OR = 2.09, \chi^2(1, N = 3200) = 25.06, p < .001$). As for ToM question type, reality questions elicited the most accurate responses compared to other question types, ($OR = 1.77, \chi^2(4, N = 3200) = 335.57, p < .001$).

As shown in Fig. 5, our examination of the models' explanations showed that GPT-4 performed well in correct reasoning and going beyond the scope of the question when providing answers in the Unexpected Contents Task in English, while GPT-3.5's capabilities were limited. Similarly, in Turkish, GPT-4 was more likely to provide correct reasoning and beyond scope explanations than GPT-3.5 in the Unexpected Contents Tasks. However, the overall correctness of the models was higher in English than in Turkish. In addition, the Unexpected Location Transfer Task appeared to be challenging for both models in English and Turkish.

In Study 2, we also zoomed in on false belief verbs and examined whether verb type was associated with the accuracy of responses (Fig. 6). We observed significant differences across false belief verb types in regard to the accuracy of responses, $\chi^2(3) = 39.81, p < 0.001$. Pairwise comparisons using Holm correction showed that Turkish and English were significantly different from each other in regular verbs, $p < .05$ and Turkish yielded more accurate responses than English. But there were no differences in false belief verbs *san-* vs. *zannet-* for eliciting accurate responses. We conducted a similar analysis for correct reasoning in explanations and found that verb type was significantly associated with accuracy of reasoning in explanations as well $\chi^2(3) = 27.86, p < 0.001$. Pairwise comparisons showed that regular English and Turkish belief verbs were not different from each other, $p > .05$. Turkish false belief verb *san-* yielded more accurate explanations compared to English and Turkish regular verbs, $p < .001$. There was no significant difference between Turkish false belief verbs *san-* and *zannet-* $p > .05$. Of all responses, only 4.5% went beyond scope in English and only 2% went beyond scope in Turkish. Therefore, in this study we did not conduct any statistical analysis on the beyond scope category.

5.3.2. Log probability results

The log probabilities in Study 2 yielded similar results to Study 1 (see Fig. 7). Overall, GPT-4 had higher confidence in belief prompts than GPT-3.5. For Turkish belief prompts, there was a notable decrease in GPT-4's confidence. While GPT models showed moderate to high confidence in reality check questions, their confidence was low in action questions. For instance, GPT-4 either showed perfect confidence (1) or lack of confidence (0), which suggests that the performance in these types of questions is far from robust at the moment. Similar to earlier findings and text sampling results, the models' confidence was lower in the Unexpected Location Transfer task than in the Unexpected Contents Task.

In sum, similar to Study 1, both text sampling and log probability results indicated that GPT-4 performed better than GPT-3.5, particularly when answering false belief questions. GPT models performed better in Unexpected Contents Tasks than in Unexpected Location Transfer Tasks. In response to action questions that ask what would the character do next in the scenario, there was a very inconsistent pattern in GPT models' responses (100% accuracy or no accuracy at all). In text sampling results, while performance appeared to be better in Turkish, no such difference was found in log probability analysis. Text sampling analysis also showed that GPT-4 was more likely to provide correct reasoning and go beyond the scope when responding to open-ended ToM questions than GPT-3.5. When Study 1 and Study 2 are compared, it is also important to note that asking the questions within a continuous chat session improved the models' performances.

6. Study 3: second-order ToM understanding of GPT models in English and Turkish

Our findings in Studies 1 and 2 replicated Ullman's findings, showing that variations in the scenarios can lead to failures in ToM performance in these models despite slight improvements in performances when questions were presented within the same chat session. GPT-4 outperformed GPT-3.5 in first-order tasks. Besides, to our surprise, GPT models performed better in Turkish than in

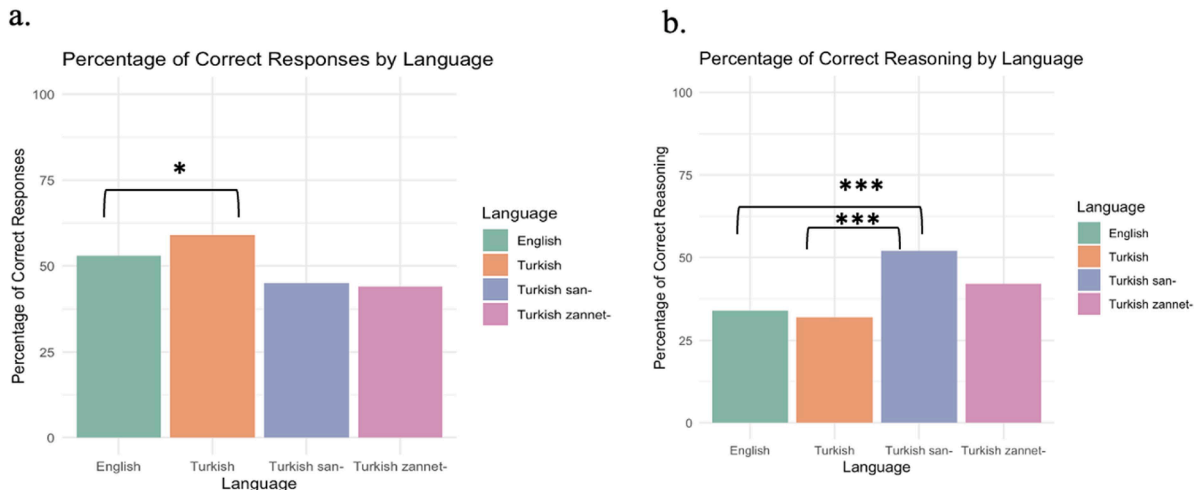


Fig. 6. Percentage of correct responses and explanations across different verb types in Study 2.

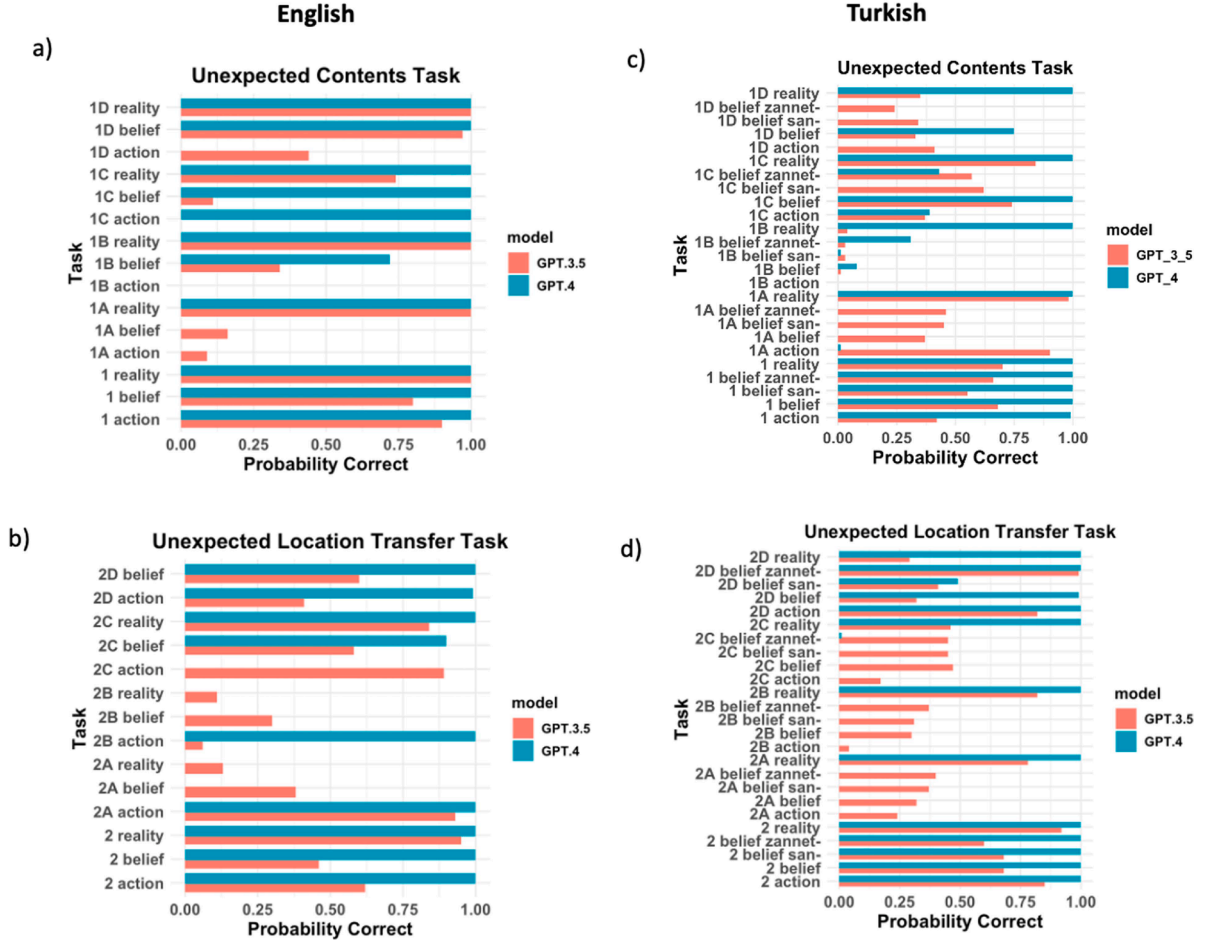


Fig. 7. Probability of correctness in the ToM tasks across language and GPT models in Study 2.

English. In Study 3, we extended our exploration of GPT models' ToM understanding by administering second-order tasks, which include recursive belief and knowledge structures. In the process, we aim to understand the extent to which GPT can and cannot reason about these tasks in English and Turkish, how many perspectives it can track within a ToM scenario, and whether it can explain the responses plausibly/give high-quality responses.

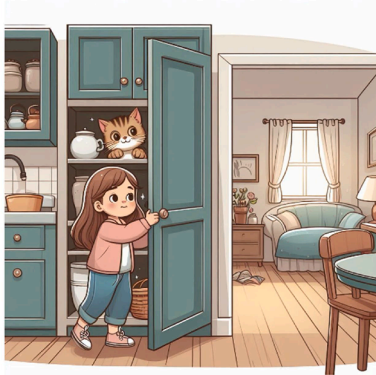
6.1. Materials and prompts

Second-order ToM tasks were adapted from Wimmer and Perner (1985) and scripted again the purposes of this study. After adapting them, we showed these scenarios to approximately 10 Turkish-speaking adults to validate them. Adults can reliably pass the second-order ToM questions given after these scenarios. We used four scenarios in total: 3A) *Bazaar*, 3B) *Kitten*, 4A) *Pie*, and 4B) *Balloonman*. The scenarios either included mistaken beliefs about content (items sold in a bazaar or hidden contents of a gift) or location (the placement of a pie or the location of a balloonman) (see Fig. 8 for visualizations by DALL-E). Similar to Study 2, second-order ToM prompts/questions were asked successively within the same chat session. Below, we provide short information about each story and ToM prompts/questions.

In the 3A) *Bazaar* scenario, there are two characters, Karen and Bea, and there is a bazaar at school. Karen goes to the bazaar to buy some chocolate cake. But then a third character, Bruce comes from the bazaar and tells Bea that there is no chocolate cake at the bazaar, and that rather honey buns are the only item being sold. In reality, there are cinnamon cookies as well as honey buns available at the bazaar and in the end, Karen buys some cinnamon cookies. After telling this scenario, we ask the following prompts/questions:

1. According to Karen, Bea believes that Karen bought some {RESPONSE} (second-order belief).
2. Why does Karen believe that Bea believes so? (second-order explanation)
3. Bea believes that Karen bought some {RESPONSE} (first-order belief).
4. Does Karen know that Bruce told Bea that there isn't any chocolate cake at the bazaar? (knowledge check)
5. Karen bought some {RESPONSE} from the bazaar (reality check).

3A) Bazaar



3B) Kitten



4A) Pie



4B) Balloonman



Fig. 8. Visualizations of second-order ToM scenarios in Study 3 generated by DALL-E 3.

6. Bruce believes that {RESPONSE} is being sold at the bazaar (first-order belief).

In the 3B) *Kitten* scenario, a character called Rebecca has her birthday approaching and wishes for a cat as a present but her mother tells her that she got her a toy instead of a cat. Later, Rebecca discovers a cat hidden in the kitchen cupboard and believes her mother has got her a cat. After Rebecca leaves, her father enters the kitchen and finds the cat and believes that it is their neighbor's and brings it to them. For this scenario, we ask the following prompts/questions:

1. Rebecca thinks that her mother got her a {RESPONSE} for her birthday (first-order thought).
2. Why does Rebecca think so? (first-order explanation)
3. Rebecca's mother got Rebecca a {RESPONSE} for her birthday (reality check).
4. According to Rebecca's mother, Rebecca believes that she will get a {RESPONSE} for her birthday (second-order thought)
5. Does Rebecca's mother know that Rebecca saw the cat? (knowledge-check)
6. What does Rebecca's father think when he sees the cat? (first-order thought)

In the 4A) *Pie* scenario, there are two characters in a classroom: Jane and Dwight. Jane has a strawberry pie that she does not share with Dwight which she puts into the cupboard. After Jane leaves, Dwight takes the pie from the cupboard and puts it inside a basket. Jane sees Dwight's actions through the window, which Dwight is not aware of. When neither Dwight nor Jane are in the room, their teacher comes in and puts the pie inside a drawer. For this scenario, we ask the following prompts/questions:

1. Dwight believes that Jane will look for the pie inside of the {RESPONSE} (second-order belief).
2. Why does Dwight believe so? (second-order explanation)
3. Does Jane know that Dwight put the pie in the basket? (knowledge check)
4. Does Dwight know that Jane saw him put the pie in the basket? (knowledge check)
5. The pie is actually in the {RESPONSE}. (reality check)
6. Jane believes that the pie is in the {RESPONSE}. (first-order belief)

7. Jane will look for the pie in the {RESPONSE}. (first-order action)
8. Dwight believes that the pie is in the {RESPONSE}. (first-order belief)

In the 4B) *Balloonman* scenario, there are two characters, Catherine and John playing at the park. A balloonman comes to the park from whom Catherine wants to buy a balloon. Catherine doesn't have any money on her, which means she has to go home to get some from her mother. The balloonman assures her and John that he will be at the park the entire afternoon. However, after a while, the balloonman tells John that since the park is empty, he will go to the school to sell balloons. On his way, he runs into Catherine and informs her as well. Afterwards, Balloonman changes his mind and goes to the shopping mall instead. John goes to Catherine's house but cannot find her. Catherine's mother tells him that she left to buy some balloons. For this scenario, we ask the following prompts/questions:

1. John thinks that Catherine went to the {RESPONSE} to buy some balloons. (second-order thought)
2. Why does John think that she went there instead of somewhere else? (second-order explanation)
3. Does John know that the balloon man told Catherine where he would be going? (knowledge check)
4. The balloon man is at the {RESPONSE}. (reality check)
5. Catherine thinks that the balloon man is at the {RESPONSE}. (first-order thought)
6. Catherine will look for the balloon man at the {RESPONSE}. (first-order action)
7. John thinks that the balloon man is at the {RESPONSE}. (first-order thought)

6.2. Data preparation and coding

We used the same coding scheme for this study as we have used in Studies 1 and 2. This study had 4 scenarios: 2 second-order beliefs, 2 second-order thoughts, and 2 explanation prompts/questions. Besides, this study included reality and knowledge check questions in each scenario as well as first-order false belief questions from the perspectives of different characters. We collapsed the

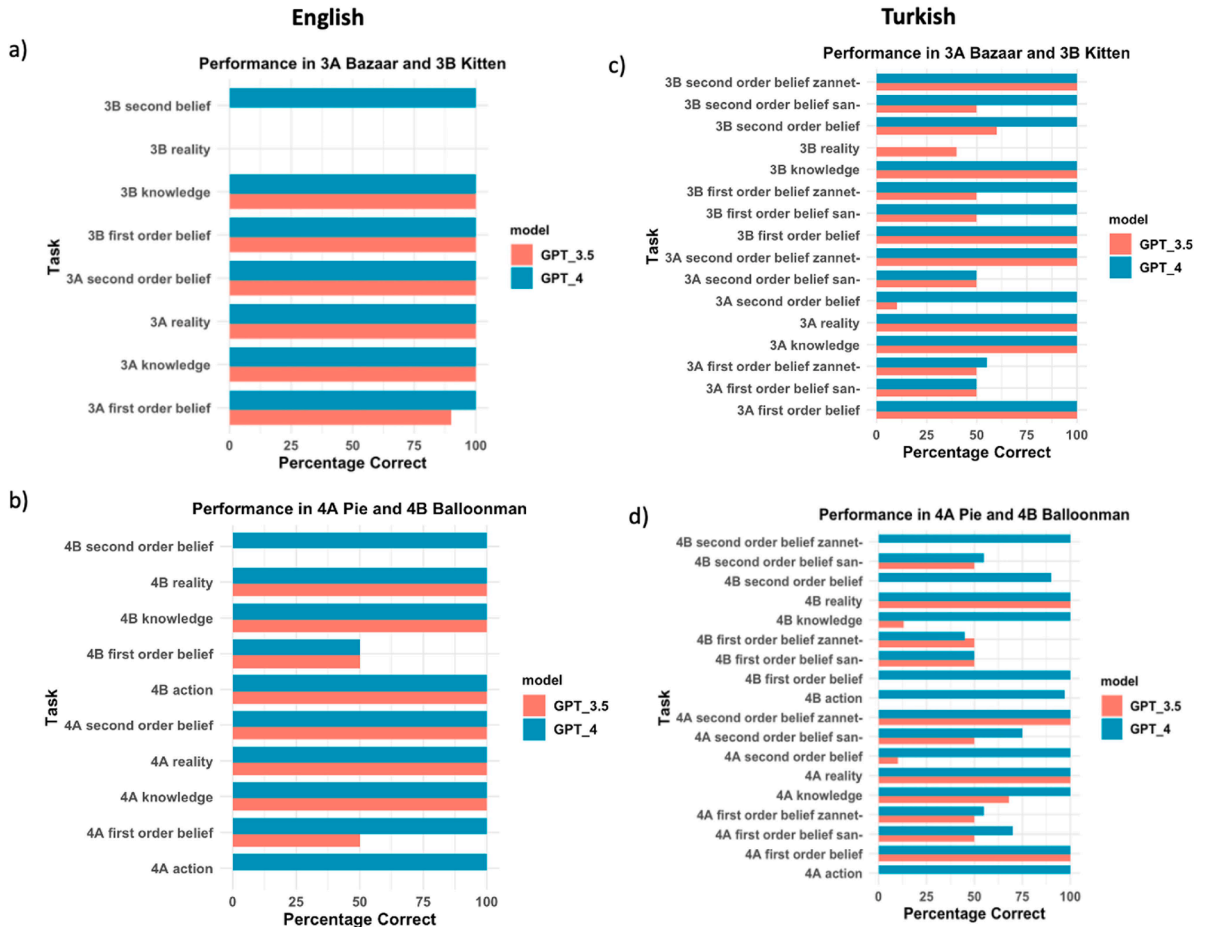


Fig. 9. Accuracy of responses in second-order ToM tasks across GPT models in Study 3.

percentage of correct responses to first-order false belief questions in the Figures. For explanation questions, we again used manual coding for 1) accuracy, 2) correct reasoning and 3) going beyond scope. As we calculated interrater reliability for all responses, the percentage agreement and Cohen's kappa values provided above in Study 1 and Study 2 included responses from Study 3 as well, indicating high agreement between coders.

6.3. Results

6.3.1. Text sampling results

In this study, we elicited a total of 2160 responses. We had the highest accuracy percentage: 1629 (75%) correct responses in all question types (see Fig. 9). For this study, we again built logistic regression models. First, we used the glmer function in R and took chat session ID as random effects since we asked all the questions within the same chat session (nested data in repeated design). However, the boundary fit was singular suggesting that there is almost no variation across sessions and random structure may not be necessary in the model. Thus, we built a simpler model by excluding random effects. We examined the effects of Language (English vs. Turkish), Story (Bazaar, Kitten, Pie, Balloonman), Model (GPT-3.5 vs. GPT-4) and ToM Question Type (Action, Belief, Explanation, Knowledge, Reality, Thought) on the accuracy of responses. Results showed that there were differences in GPT's performance across different scenarios ($\chi^2(1, N = 2160) = 20.81, p < .001$). Balloonman and Kitten were the most difficult scenarios, the scenarios had higher likelihood of correct responses (Kitten: $OR = .57$, Pie: $OR = 1.76$, Bazaar: $OR = .207$). Contrary to our findings in Study 1 and 2, in these second-order scenarios, GPT models were more likely to provide correct responses in English than in Turkish, ($OR = .57, \chi^2(1, N = 2160) = 17.20, p < .001$). GPT 4 provided more accurate responses to second order ToM questions than GPT 3.5 ($OR = 4.76, \chi^2(1, N = 2160) = 199.62, p < .001$). ToM question type was also a significant predictor of correct responses, ($\chi^2(5, N = 2160) = 114.26, p < .001$). The odd ratios for different question types indicate that compared to Thought questions, Action ($OR = .62$), Belief ($OR = 1.55$), Knowledge ($OR = 1.18$, Reality ($OR = .61$), and Explanation ($OR = 1.97$) questions had a higher likelihood of eliciting correct responses.

When we examined the responses to open-ended explanation questions, we found that GPT-4 generally outperformed GPT-3.5 in providing explanations that follow correct reasoning. Overall, GPT models were better in providing explanations in English than in Turkish. Moreover, GPT models provided better explanations in scenarios 3A and 3B which are related to understanding of Unexpected Contents than 4A and 4B which are related to Unexpected Location Transfer (see Fig. 10).

Verb type was significantly associated with the accuracy of responses, $\chi^2(3) = 20.3, p < 0.001$. Pairwise comparisons using Holm correction showed that English had higher accuracy in responses and reasoning than Turkish and Turkish with false belief verbs, $ps < .05$. Similarly, verb type was significantly associated with accuracy of reasoning in explanations as well $\chi^2(3) = 55.54, p < 0.001$. Pairwise comparisons showed that Turkish *san-* and *zannet-* were not different from each other, but English had much higher correct reasoning than other verb types $ps < .001$ (see Fig. 11).

6.3.2. Log probability results

Findings showed that GPT-4 had generally higher log probabilities of correct responses in second-order ToM tasks than GPT-3.5, indicating GPT-4 had higher confidence in these complex reasoning tasks than GPT-3.5. Contrary to what we observed in the text sampling analysis, log probabilities were more consistent and slightly higher in English than in Turkish (see Fig. 12). We did not observe any systematic patterns of change in the performance when Turkish false belief verbs were used. These findings suggest that GPT models have not yet mastered ToM tasks to become sensitive to linguistic nuances in differences, unlike young children who can make use of such linguistic cues. Similar to Studies 1 and 2, Unexpected Location Transfer Tasks (4A and 4B) posed more challenges to

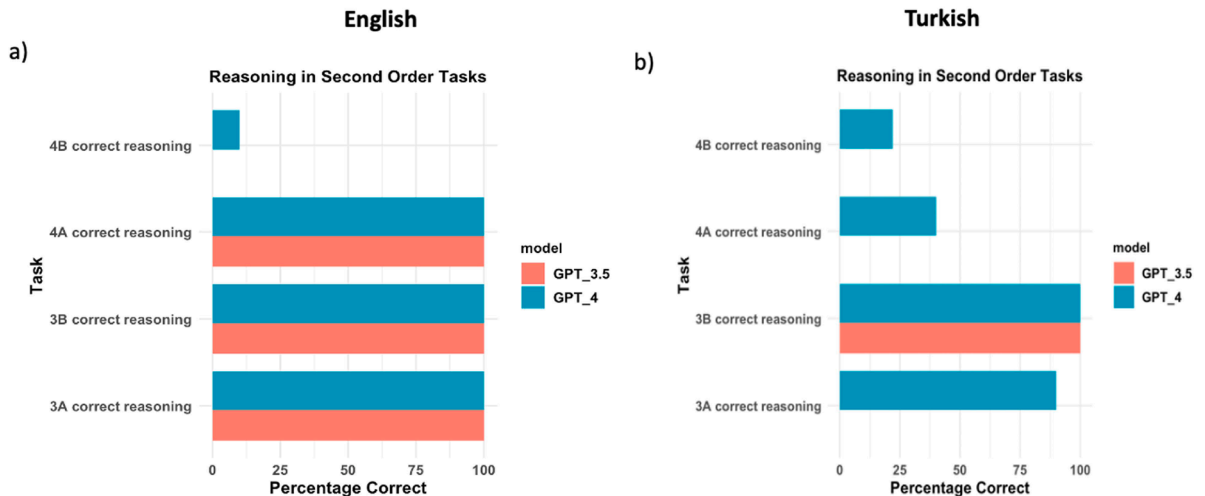


Fig. 10. Quality of responses to second-order explanation questions in Study 3.

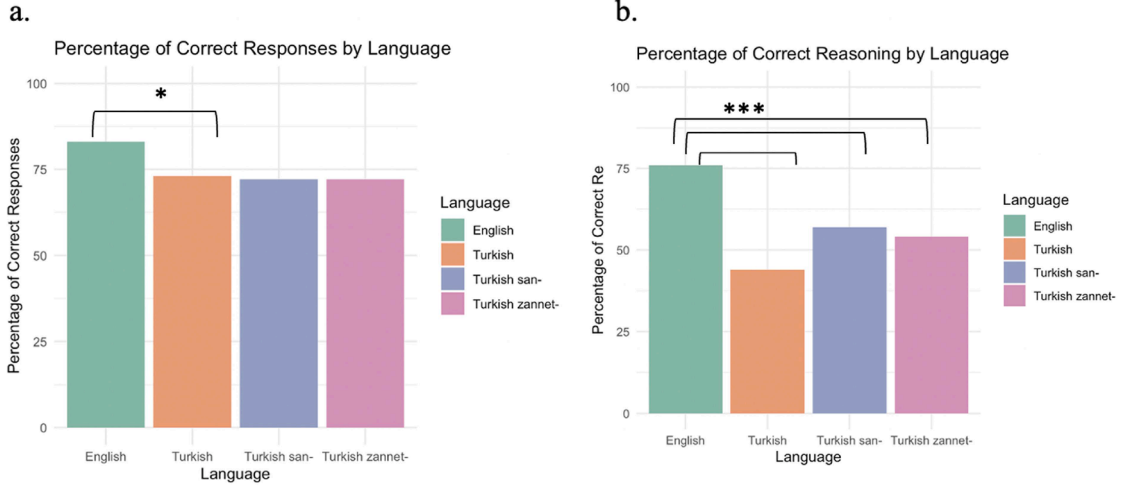


Fig. 11. Percentage of correct responses and explanations across different verb types in Study 3.

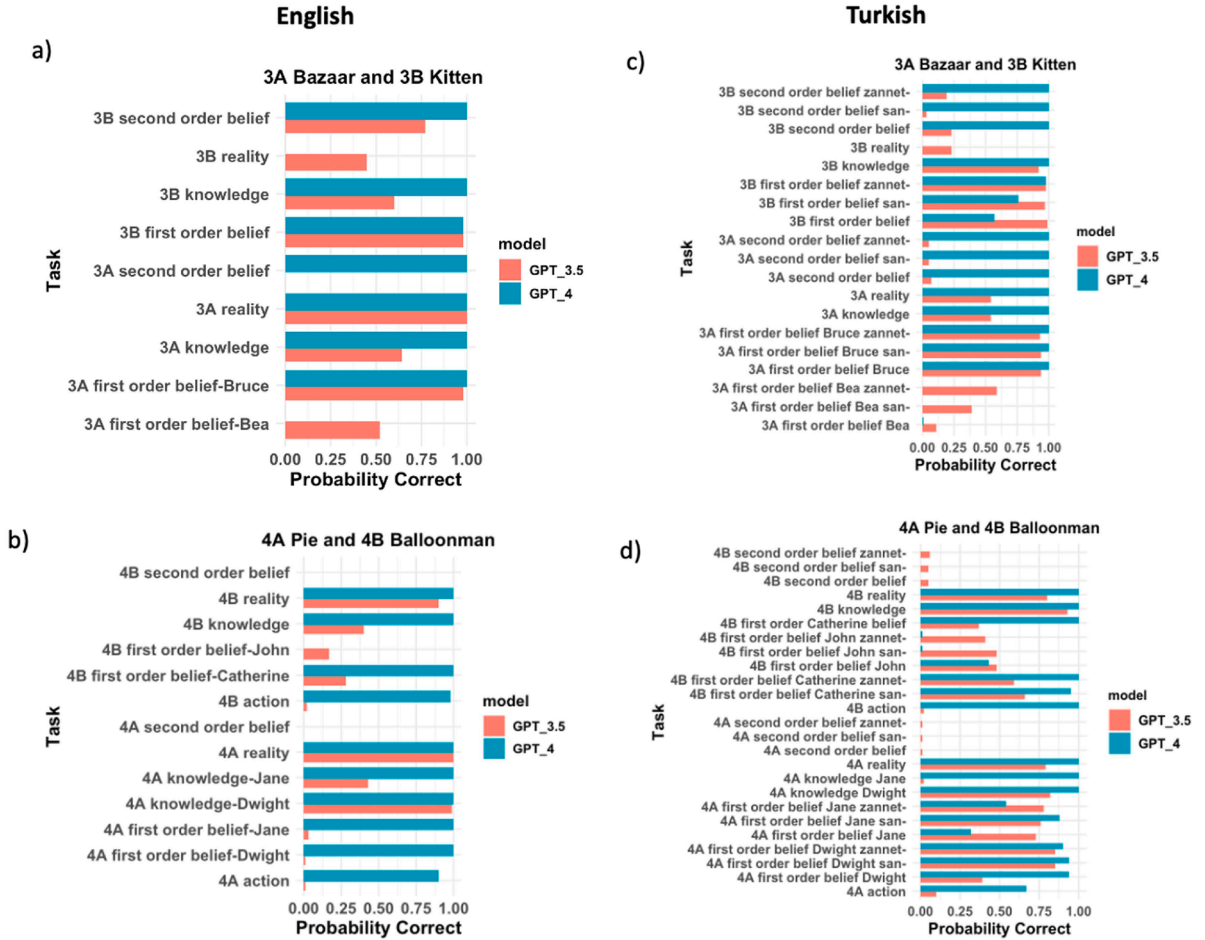


Fig. 12. Probability of correctness in the ToM tasks across language and GPT models in Study 3.

GPT models than Unexpected Contents Tasks (3A and 3B). Overall, we have also seen lower log probabilities in action questions. This is a pattern we observed in Studies 1 and 2 as well, suggesting that predicting a character's action based on belief presents a special difficulty for GPT models. Contrary to text sampling results, we have not observed any systematic differences related to the use of false

belief verbs in the questions/prompts in Turkish. (See Appendix B detailed examination of the probabilities per prompt by language and scenario).

To summarize, in both text sampling and log probability analyses, GPT models performed better in English than in Turkish in the second-order ToM tasks. GPT-4 yielded higher confidence and more accuracy in the second-order tasks than GPT-3.5. There was high variability across question types (action, belief, thought, knowledge). Similar to Studies 1 and 2, GPT models had difficulty providing correct responses to action questions in particular. These findings suggest that GPT models still face challenges when responding to ToM questions, especially with linguistic nuances and action-based prediction prompts. For open-ended explanation prompts, we found that GPT models performed better when dealing with linguistic structures in English than in Turkish, especially in complex reasoning scenarios.

7. Discussion

Across three studies, we investigated the ToM capabilities of GPT-3.5 and GPT-4 models in two different languages, emphasizing the significance of assessing these capabilities using tools from the developmental psychology literature to gain insight into how machines and children develop social understanding. In Study 1, we replicated Ullman's (2023) study by assessing the first-order ToM performance of GPT 3.5 GPT 4 in English and Turkish (Study 1). Second, we built upon the initial study and expanded it to include close-ended and open-ended questions and examined how GPT models respond to them within the same chat session in English and Turkish (Study 2), which allows for a more realistic assessment by simulating naturalistic conversations and different types of questions. Third, we examined the performance of GPT-3.5 and GPT-4 models in more complex second-order false belief scenarios in Turkish and English (Study 3). Consistent with Kosinski's observations, GPT-4 outperformed its predecessor, GPT-3.5; however, except for the baseline tasks, GPT-4's accuracy showed huge fluctuations based on the task type and prompts. Supporting Ullman's findings, we found that GPT models generally respond accurately to baseline ToM prompts/questions; however, their performance drops dramatically with any variations in the scenarios and questions, suggesting that these models cannot reliably pass ToM tasks yet. Our examination of the performance in two languages showed that although models performed similarly in both languages, there were some linguistic nuances in tasks, and models need more flexibility in different languages. Similar to earlier arguments regarding the models' capability to only imitate statistical patterns without genuine comprehension, we found that GPT-3.5 and GPT-4 cannot perform as well as young children who can pass first-order ToM tasks around 4-5 years of age and second-order tasks around 6 years of age yet (e.g., Wellman et al., 2000; Wimmer and Perner, 1985). Our findings contribute to the ongoing discussion about the current capabilities, limitations, and potential of LLMs in ToM understanding. Below, we will discuss our specific findings in detail.

Across three studies, both text sampling and log probability results showed that GPT-4 was more accurate than the earlier model GPT-3.5 in responding to ToM questions and providing explanations (Kosinski, 2023; Bubeck et al., 2023; Open AI Research, 2023). However, as mentioned above, slight variations in the scenarios, for instance, asking the model to perceive and respond according to what the concept of transparency entails or trusting testimony means, decreased the performance of even GPT-4 significantly (Ullman, 2023). A major contribution of our study is asking open-ended explanation questions to GPT models and assessing whether these explanations have accurate reasoning or not. This analysis showed that although GPT-4 cannot respond above chance levels to close-ended ToM prompts yet, it can provide reasonable explanations for false-belief-based behavior and can even go beyond scope by speculating about possible causes. For instance, in the Unexpected Contents Task in which the bag was mislabeled, GPT-4 provides explanations, saying that the character can smell the popcorn or shake the bag to infer the contents from the sound. These speculative explanations indicate that the model can provide some contextually relevant reasoning based on its training data but are still far from consistent and require further training and refinement.

Despite some variability across prompts and scenarios, both GPT-3.5 and GPT-4 performed better in general in responding to English scenarios and prompts than Turkish scenarios and prompts. This difference can be explained by the availability of large training data in English, which allows the models to have a better understanding of ToM prompts (Penedo et al., 2023). On the other hand, Turkish with its smaller training data compared to English and different morphological and syntactic structure poses a greater challenge to the models even though the existence of false belief verbs that directly denote the presence of a false belief in a given scenario (Ofaşer and Saraçlar, 2018; Penedo et al., 2023). This finding highlights the importance of linguistic nuances and training data for LLMs to show ToM understanding. Thus, in the future, it is important to diversify training data to achieve higher flexibility and accuracy in different languages.

Both GPT models showed great success in baseline ToM tasks. However, they consistently showed lower performance in Expected Location Transfer Tasks than in Unexpected Contents Tasks. Considering children's performance, as shown in developmental psychology literature, it is known that children show a progression in their understanding of diverse desires, beliefs, knowledge, and false beliefs. However, their performances do not particularly differ in Unexpected Contents and Unexpected Locations Tasks (e.g., Wellman et al., 2001; Wellman and Liu, 2004). There are some findings in the literature suggesting that children could perform better in Unexpected Location Transfer Tasks than in Unexpected Contents Tasks as Contents tasks (e.g., Holmes et al., 1996). The opposite pattern of performance in GPT models suggests that LLMs learn and apply knowledge differently than children when responding to these false belief tasks (Kosoy et al., 2023). It is possible that certain characteristics of the Location Transfer Task could have challenged LLMs. When children master ToM tasks, they learn via their first-hand experiences, active exploration of the world, perceiving object-space relations, and communication with others. On the other hand, LLMs have language proficiency, but they lack first-hand experience. Thus, this lack of experience may lead LLMs to just imitate social understanding in a limited fashion (Yui et al., 2023). In future studies, the mechanisms behind this pattern can be tested with more advanced GPT models, including multimodal features like videos and pictures, as they can become more attuned to navigating spatial layouts and tracking object locations.

GPT models' performance was very low in Ullman's variations, but the performance did not drop as much in second-order ToM tasks. These findings suggest that the models can mimic statistical patterns very well and can do well in tasks that require verbal proficiency, although the tasks have complex embedded false beliefs. However, when it comes to task variations that require some real-world experience, such as knowing the concept of transparency, being literate, change in spatial relations, or the notion of the testimony of a friend, the models failed dramatically. Thus, achieving a full-fledged ToM understanding requires not only high language proficiency and perspective-taking skills but also real-world conceptual understanding (Yui et al., 2023).

Receiving the prompts/questions within the same chat session, like a prolonged conversation, improved the GPT models' performance. This finding suggests that Chat Completion API, which allows for ongoing, contextual conversation, eases the interaction between humans and LLMs by enhancing LLMs' performance in ToM. However, there is still a need for improvement to enhance coherence within a conversational context with a human and LLM.

Our dual-coded manual analyses of GPT models' explanations in Studies 2 and 3 allowed us to gain insight into the reasoning steps of GPT models, which utilize zero-shot learning (Penedo et al., 2022; Wolfram, 2023). We assessed GPT models' current capacity and limitations to replicate human-like thinking processes and speculate when confronted with uncertainty (e.g., uninformative label scenario in first-order false belief tasks). We also found out that when speculating about explanations for a judgment, GPT models, particularly GPT-4, can go beyond the scope of current scenarios and make inferences by integrating background knowledge and, in a sense, imagination. This examination of explanations paves the way for understanding the *Explainability* of AI models, especially ones that rely on complex deep learning processes. It also provides a holistic understanding of the strengths and weaknesses of the GPT models by offering a model of their engagement in linguistic and cognitive processes (Aher et al., 2023; Bubeck et al., 2023; Van Dijk et al., 2023).

Simultaneously, though, we should be aware of our human tendency to anthropomorphize these AI models when we interpret their responses, as showing a certain level of performance and having true comprehension are two different issues. As we know from the famous Chinese Room Experiment, machines may behave as if they have the skill at the practical level and make us believe that they have human-like reasoning. While LLM models may communicate, indicating some degree of proficiency, this does not necessarily mean that they have an understanding. On a practical level, this phenomenon reminds us of the proverb "fake it till you make it," suggesting that LLMs may just be mimicking statistical patterns without real human-like comprehension today, but with continued advancements in the field, possibly one day, they can reach a level indistinguishable from humans (Kosinski, 2023; Ullman, 2023, Yui et al., 2023).

While this study provided valuable insights into the ToM capabilities of GPT models in English and Turkish, we acknowledge several limitations in our experimental design. Specifically, we set the system message as "You are a helpful assistant" and the temperature parameter to 0 to ensure consistent and reproducible results; however, this decision could have limited the diversity of responses we received from the GPT models and our understanding of the how these models may behave in more dynamic interaction settings. Thus, it remains an open question how varying the persona (e.g., human participant, 5-year-old child, female, male) can impact the models' interactions with humans. Future research could investigate the adaptability of LLMs by designing experiments in which the model assumes different roles in specific contexts. This research could provide deeper insights into the flexibility of ToM skills in LLMs in meeting the demands and expectations of humans during interactions. Likewise, varying the temperature parameter was beyond the scope of this study, but researchers could explore the generative capabilities of the models by heightening the temperature and observing the balance between coherence and imaginativeness in artificial cognition.

Moreover, it is important to acknowledge that the specific theory of mind tests used in this study are versions of verbal scenarios presented to children and adults in the developmental psychology literature. While they still provide valuable insights into the capabilities and restrictions of LLMs, their applicability to LLMs as tests of evidence of theory of mind may be limited. Specifically, these tests may not fully capture the complexities of having a theory of mind understanding in real-life social interactions. Thus, future research should adopt more comprehensive and ideally multimodal tests to address these limitations.

8. Conclusion

Our study contributes to the existing literature by assessing ToM understanding of GPT models in two languages (English and Turkish) and providing findings related to differences in linguistic characteristics of languages and cognitive demandingness of ToM tasks. Our methodology and approach provide important insights into LLMs' current capabilities and ways of reasoning and establish a standard for future research in this area. As AI models continue to develop, they will continue to interact with humans. Having ToM understanding is necessary for smooth social interaction; since our daily lives and social interactions with others are interwoven with different mental states, ToM understanding has utmost importance for adult social life as we experience misunderstandings or communication failures in our everyday interactions due to our or others' false beliefs. Research in the field of psychology also indicates that practice enhances ToM skills; reading literary fiction with different characters, mental states, and events improves ToM understanding in adults (Kidd and Castano, 2013). Similarly, exposing GPT models to high-quality input from fictional narratives could provide "life-like" situations to the models and potentially improve their ToM understanding (Bruner, 1986).

Moreover, children do not develop ToM understanding instantly, it happens over several years and by being embedded in social interaction settings. GPT models' responses to Ullman's ToM task variations make us think that GPT is like an agent who has been given a lot of information about the world in a dark room. It has knowledge, but it lacks pragmatics. Human-level pragmatic skills, ToM understanding, empathy and emotional understanding, and intrinsic motivation for exploration and learning from others require being exposed to social interactions with humans (Yiu et al., 2023; Du et al., 2023). Training these AI models among humans with social experience and meaningful interactions can lead to improvements in understanding mental states; one day we may turn to them for

advice in complex real-life situations as socially intelligent agents.

If LLMs can achieve this milestone, we believe that these tools can help individuals who struggle with ToM understanding (e.g., individuals with autism spectrum disorder) in complex social contexts. While we recognize that these models may not be as flexible and intuitive as humans in detecting intentions, they are developing a capacity to infer mental states from intricate situations. This observation highlights the potential applications of these tools in social life. The use of GPT models in complex social contexts may provide significant assistance and improvement of social cognition for people who face challenges in these areas and thereby improve their quality of social relationships.

Moving forward, exploring multimodal ToM tasks, which could be integrated into the latest versions of GPT with vision capabilities and other AI models (e.g., Jin et al., 2024), and changing the persona of the language model by changing the system message from “You are a helpful assistant” to more detailed descriptions are promising future avenues for research. Moreover, future research should focus on multilingual contexts and examine how specific language contexts, including word choice, sentence structure, grammatical nuances, and the use of language in context, may impact the ToM performance of LLMs. Such studies can lead to a more profound understanding of the linguistic and cognitive processes that are needed for understanding other people’s mental states. As advances in AI occur, it also becomes crucial to have accurate input/data and be sensitive to different cultural and linguistic characteristics. The capacity and effectiveness of AI models to reliably understand and interact with humans depends on their inclusiveness and diversity. Thus, future research should contribute to finding ways to enable culturally competent and empathetic human-machine interactions.

CRedit authorship contribution statement

Burcu Ünlütürk: Conceptualization, Writing – original draft, Writing – review & editing, Visualization, Methodology, Formal analysis, Data curation, Supervision. **Onur Bal:** Writing – review & editing, Visualization, Methodology, Formal analysis, Data curation.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

I have attached link to my data in the method section of the manuscript.

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at [doi:10.1016/j.csl.2024.101698](https://doi.org/10.1016/j.csl.2024.101698).

References

- Aher, G.V., Arriaga, R.I., Kalai, A.T., 2023. Using large language models to simulate multiple humans and replicate human subject studies. In: *Proceedings of the International Conference on Machine Learning*. PMLR, pp. 337–371.
- Akhtar, N., Gernsbacher, M.A., 2008. On privileging the role of gaze in infant social cognition. *Child Dev. Perspect.* 2 (2), 59–65.
- Arslan, B., Taatgen, N.A., Verbrugge, R., 2017. Five-year-olds’ systematic errors in second-order false belief tasks are due to first-order theory of mind strategy selection: a computational modeling study. *Front. Psychol.* 8, 275.
- Bartsch, K., Wellman, H.M., 1995. *Children talk about the mind*. Oxford University Press.
- Bates, D., Mächler, M., Bolker, B., Walker, S., 2015. Fitting linear mixed-effects models using lme4. *J. Stat. Softw.* 67 (1), 1–48. <https://doi.org/10.18637/jss.v067.i01>.
- Binz, M., Schulz, E., 2023. Using cognitive psychology to understand GPT-3. *Proc. Natl. Acad. Sci.* 120 (6), e2218523120.
- Bostrom, N., 2014. *Superintelligence: Paths, Dangers and Strategies*. Oxford University Press.
- Bruner, J.S., 1986. *Actual Minds, Possible Worlds*. Harvard University Press.
- Bubeck, S., Chandrasekaran, V., Eldan, R., Gehrke, J., Horvitz, E., Kamar, E., & Zhang, Y. (2023). Sparks of artificial general intelligence: Early experiments with gpt-4. *arXiv preprint arXiv:2303.12712*.
- Call, J., Tomasello, M., 2008. Does the chimpanzee have a theory of mind? 30 years later. *Trends Cogn. Sci.* 12 (5), 187–192.
- Carpenter, M., Nagell, K., Tomasello, M., Butterworth, G., & Moore, C. (1998). Social cognition, joint attention, and communicative competence from 9 to 15 months of age. *Monographs of the society for research in child development*, 1-174.
- Carpendale, J.I., Lewis, C., 2004. Constructing an understanding of mind: the development of children’s social understanding within social interaction. *Behav. Brain Sci.* 27 (1), 79–96.
- Coda-Forno, J., Witte, K., Jagadish, A. K., Binz, M., Akata, Z., & Schulz, E. (2023). Inducing anxiety in large language models increases exploration and bias. *arXiv preprint arXiv:2304.11111*.
- Cooper, R.P., Aslin, R.N., 1989. The language environment of the young infant: implications for early perceptual development. *Can. J. Psychol. Rev. Can. Psychol.* 43 (2), 247.
- Du, Y., Kosoy, E., Dayan, A., Rufova, M., Abbeel, P., Gopnik, A., 2023. What can AI learn from human exploration? Intrinsically-motivated humans and agents in open-world exploration. In: *Proceedings of the NeurIPS 2023 workshop: Information-Theoretic Principles in Cognitive Systems*.
- Fox, J., Weisberg, S., 2019. *An R companion to applied regression*, 3rd ed. Sage, Thousand Oaks CA <https://socialsciences.mcmaster.ca/jfox/Books/Companion/>.

- Gopnik, A. (2003). The theory theory as an alternative to the innateness hypothesis. *Chomsky and his critics*, 238-254.
- Gopnik, A., Wellman, H.M., 1994. The theory theory. In: *Proceedings of the An Earlier Version of this Chapter was Presented at the Society for Research in Child Development Meeting*. Cambridge University Press, p. 1991.
- Göksel, A., & Kerslake, C. (2004). *Turkish: A comprehensive grammar*. Routledge.
- Holmes, H.A., Black, C., Miller, S.A., 1996. A cross-task comparison of false belief understanding in a Head Start population. *J. Exp. Child Psychol.* 63 (2), 263-285. <https://doi.org/10.1006/jecp.1996.0050>.
- Hu, J., Levy, R., 2023. Prompting is not a substitute for probability measurements in large language models. In: *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 5040-5060.
- Jiang, G., Xu, M., Zhu, S. C., Han, W., Zhang, C., & Zhu, Y. (2022). Mpi: Evaluating and inducing personality in pre-trained language models. *arXiv preprint arXiv:2206.07550*.
- Jin, C., Wu, Y., Cao, J., Xiang, J., Kuo, Y. L., Hu, Z., & Shu, T. (2024). Mmtom-qa: Multimodal theory of mind question answering. *arXiv preprint arXiv:2401.08743*.
- Ketrez, F.N., 2012. *A Student Grammar of Turkish*. Cambridge University Press.
- Kidd, D.C., Castano, E., 2013. Reading literary fiction improves theory of mind. *Science* 342 (6156), 377-380.
- Kosinski, M. (2023). Theory of mind may have spontaneously emerged in large language models. *arXiv preprint arXiv:2302.02083*.
- Kosoy, E., Reagan, E. R., Lai, L., Gopnik, A., & Cobb, D. K. (2023). Comparing Machines and Children: Using Developmental Psychology Experiments to Assess the Strengths and Weaknesses of LaMDA Responses. *arXiv preprint arXiv:2305.11243*.
- Lee, K., Olson, D.R., Torrance, N., 1999. Chinese children's understanding of false beliefs: the role of language. *J. Child Lang.* 26 (1), 1-21.
- Matsui, T., Rakoczy, H., Miura, Y., Tomaselli, M., 2009. Understanding of speaker certainty and false-belief reasoning: a comparison of Japanese and German preschoolers. *Dev. Sci.* 12 (4), 602-613.
- Naito, M., Koyama, K., 2006. The development of false-belief understanding in Japanese children: delay and difference? *Int. J. Behav. Dev.* 30 (4), 290-304.
- Oflazer, K., Saraçlar, M., 2018. *Turkish natural language processing*. Springer International Publishing.
- OpenAI. 2024. GPT-4. Retrieved from <https://openai.com/research/gpt-4>.
- OpenAI Community. (2023, July 6). Logprobs are missing from the Chat endpoints. OpenAI Developer Forum. Retrieved from <https://community.openai.com/t/logprobs-are-missing-from-the-chat-endpoints/289514>.
- Penedo, G., Malartic, Q., Hesslow, D., Cojocaru, R., Cappelli, A., Alobeidli, H., & Launay, J. (2023). The RefinedWeb dataset for Falcon LLM: outperforming curated corpora with web data, and web data only. *arXiv preprint arXiv:2306.01116*.
- Perner, J., & Howes, D. (1992). "He thinks he knows": And more developmental evidence against the simulation (role taking) theory. *Mind Language*.
- Perner, J., Wimmer, H., 1985. John thinks that Mary thinks that..." attribution of second-order beliefs by 5-to 10-year-old children. *J. Exp. Child Psychol.* 39 (3), 437-471.
- Peterson, C.C., Siegal, M., 1998. Changing focus on the representational mind: deaf, autistic and normal children's concepts of false photos, false drawings and false beliefs. *Br. J. Dev. Psychol.* 16 (3), 301-320.
- Ruffman, T., 2023. Belief it or not: How children construct a theory of mind. *Child Dev. Perspect.*
- Selcuk, B., Brink, K.A., Ekerim, M., Wellman, H.M., 2018. Sequence of theory-of-mind acquisition in Turkish children from diverse social backgrounds. *Infant Child Dev.* 27 (4), e2098.
- Shahaeian, A., Nielsen, M., Peterson, C.C., Slaughter, V., 2014. Cultural and family influences on children's theory of mind development: A comparison of Australian and Iranian school-age children. *J. Cross-Cult. Psychol.* 45 (4), 555-568.
- Shatz, M., Diesendruck, G., Martinez-Beck, I., Akar, D., 2003. The influence of language and socioeconomic status on children's understanding of false belief. *Dev. Psychol.* 39 (4), 717.
- Tardif, T., Wellman, H.M., Cheung, K.M., 2004. False belief understanding in Cantonese-speaking children. *J. Child Lang.* 31 (4), 779-800.
- Turati, C., Macchi Cassia, V., Simion, F., Leo, I., 2006. Newborns' face recognition: role of inner and outer facial features. *Child Dev.* 77 (2), 297-311.
- Ullman, T. (2023). Large language models fail on trivial alterations to theory-of-mind tasks. *arXiv preprint arXiv:2302.08399*.
- Ünlütürk, B., Aksu-Koç, A., 2015. Developmental relations between reference to characters in narratives and theory of mind. Zeyrek, D., Sağın-Şimşek, Ç. Atas, U., and Rehbein, R. (Eds.). *Ankara Papers in Turkish and Turkic Linguistics*. Harrassowitz Verlag, Wiesbaden.
- van Duijn, M. J., van Dijk, B., Kouwenhoven, T., de Valk, W., Spruit, M. R., & van der Putten, P. (2023). Theory of Mind in Large Language Models: Examining Performance of 11 State-of-the-Art Models vs. Children Aged 7-10 on Advanced Tests. *arXiv preprint arXiv:2310.20320*.
- Vinden, P.G., 2001. Parenting attitudes and children's understanding of mind: a comparison of Korean American and Anglo-American families. *Cogn. Dev.* 16 (3), 793-809.
- Wellman, H.M., 2002. Understanding the psychological world: Developing a theory of mind. *Blackwell Handb. Child. Cogn. Dev.* 167-187.
- Wellman, H.M., Cross, D., Watson, J., 2001. Meta-analysis of theory-of-mind development: the truth about false belief. *Child Dev.* 72 (3), 655-684.
- Wellman, H.M., Liu, D., 2004. Scaling of theory-of-mind tasks. *Child Dev.* 75 (2), 523-541.
- Wellman, H.M., Fang, F., Liu, D., Zhu, L., Liu, G., 2006. Scaling of theory-of-mind understandings in Chinese children. *Psychol. Sci.* 17 (12), 1075-1081.
- Wickham, H., 2011. The split-apply-combine strategy for data analysis. *J. Stat. Softw.* 40, 1-29. <https://doi.org/10.18637/jss.v040.i01>.
- Wimmer, H., 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.
- Wolfram, S. (2023, February). What is ChatGPT doing and why does it work? Retrieved from <https://writings.stephenwolfram.com/2023/02/what-is-chatgpt-doing-and-why-does-it-work/>.
- Woodward, A.L., 1998. Infants selectively encode the goal object of an actor's reach. *Cognition* 69 (1), 1-34.
- Yiu, E., Kosoy, E., & Gopnik, A. (2023). Imitation versus Innovation: What children can do that large language and language-and-vision models cannot (yet)? *arXiv preprint arXiv:2305.07666*.



Burcu Ünlütürk is an Assistant Professor of Psychology at Yeditepe University (Turkey). She received her BA degree in English Language and Literature, and MA degree in Developmental Psychology from Boğaziçi University, and her Ph.D. in Developmental Psychology from Lehigh University. She specializes in children's language and communicative development, narrative skills, and conceptual development. Additionally, she is interested in child-computer interaction and how children acquire information from humans and technological resources (e.g., the internet and virtual dialogue systems) via question-answer exchanges.



Onur Bal was born in 1997 in Istanbul, Turkey. He graduated from Yeditepe University with two Bachelor of Arts degrees, one for English Language and Literature and the other for Psychology, in 2019 and 2020, respectively. After graduation, he worked for 2 years in the private sector as a deputy manager and English-Turkish-Chinese translator before starting his Master of Arts in Clinical Psychology at Yeditepe University in 2022. Currently, he is in his second year of studies, working at the university's hospital as an intern and writing his thesis at the same time.