Utrecht University

The Alignment Formula: Large Language Models and Humans' Decisions in a False-Belief Task

by

Mădălina Zgreabăn

A thesis submitted for the
degree of Master of Arts
in the
Faculty of Humanities

Supervisors, first and second readers:
Dr. Tejaswini Deoskar
Dr. Rick Nouwen

August 2024

Declaration of Authorship

I, Bianca Mădălina Zgreabăn, declare that this thesis titled, 'The Alignment Formula: Large Language Models and Humans' Decisions in a False-Belief Task' is written by myself. I confirm that:

- This work was conducted to complete the Research Master Degree in Linguistics at Utrecht University.
- This work has not been previously submitted to obtain another degree at another university.
- The current thesis results from my own writing and work. If I quote others, I always give the source.

Signed: 

Date: 30.08.2024

Utrecht University

Abstract

Faculty of Humanities

by Mădălina Zgreabăn

Theory of Mind (ToM) is the ability of humans to estimate the beliefs and attitudes of those around them, being highly debated regarding its origin in humans, as many argue language influences it. For example, previous studies have shown that pronouns or sentential complements lead to short and long-term improved performance of participants in ToM tasks. However, despite these improvements, it has not been previously investigated whether these linguistic constructions can also prime participants to improve their ToM abilities immediately. Secondly, ToM and its presence in Large Language Models (LLMs) is highly debated, with many arguing for or against it. Though many previous works have aimed to improve ToM in LLMs, no previous studies have examined whether pronouns or sentential complements can do so, despite the improvements mentioned above in humans. Even more, no previous study has researched what humans and LLMs pay attention to in ToM stimuli and if their attention is distributed similarly, despite recent studies showing alignment between humans and models being correlated with better model performance on various tasks.

Thus, the current thesis tests if both language models and humans can improve their performance on a new dataset for a False-belief Task when exposed to questions with sentential complements and pronouns. We use a self-paced reading task to observe which words have higher reading times and are more important for humans' answers and attribution maps to study which words weigh more in the models' predicted answers. We rank the words with the highest reading times and highest attribution scores and compare the ranks of humans and models to test their similarity. In this comparison, we also test if fine-tuning models on sentential complements or pronouns makes their rankings more similar to those of humans or if it generally improves their answers. Our results show that humans perform around 70% on our stimuli, while models perform at most around 50%. Our results also suggest that there might be a priming effect of improved performance in humans when exposed to first-person pronouns and sentential complements. Contrastively, language models do not get more correct responses when exposed to these constructions. Our results also suggest that fine-tuned language models do not get better at ToM but that they improve in traditional extractive Q&A. The current thesis also shows that the ranks of important words from models and humans are not very similar but that they do get more similar when it comes to lexical words. We also observe that fine-tuning models does not make their ranks more similar to humans.

Acknowledgements

I want first to thank my supervisors, dr. Rick Nouwen and Dr. Tejaswini Deoksar, for the help they have given me during the whole process of working and writing the current thesis. This thesis would not have been completed or complete without their guidance, patience, understanding, and trust, especially as our work together encouraged me to persist in difficult moments.

I want to thank the Utrecht University Fund, which awarded me the Bright Minds Fellowship for two years, a fund that allowed me to pursue Linguistics at UU. I feel very lucky to have pursued this degree, which broadened my horizons by exposure to many subfields of linguistics while also allowing me to pursue my interest in natural language processing. I feel extremely grateful to say that during the past two years, I have learned to think critically and creatively about linguistics and to be perseverent and courageous in my ideas and research questions. These personal achievements are the acquired abilities that I cherish the most.

Because of this, I would also like to thank all my teachers who slowly helped me grow in each of their classes. Special thanks go to my internship supervisor, Dr. Pablo Mosteiro, who trusted me that I could and helped me develop technical and analytical abilities. I also thank the professors who trusted I could use those technical abilities in projects that helped me evolve: Dr. Albert Gatt and Dr. Denis Paperno.

Finally, Linguistics at UU meant that I could meet my classmates and friends, who have positively influenced and completely changed my life, with special thanks to my dearest ones: Bruna Barbosa Louzada, Ruxandra Gîdei, Jiayi Zhou, Rishabh Suresh, and Elio Stasica. It is because of them that I feel most grateful and lucky that UU was the place where I pursued my MA.

Lastly, I would like to thank my parents for their unconditional love, support, and humourous perspective of seeing the world. Though they have not been enrolled in the program degree, they have emotionally lived every step of it like they were.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

The ability of humans to estimate other's beliefs, mental states, and attitudes is referred to as Theory of Mind (ToM) [3–8]. ToM is widely researched due to its alleged implications in common-sense reasoning [9, 10], social intelligence [11–13], among others. There are many debates concerning ToM, such as its origins, what causes and improves it [14–16], or its presence in Large Language Models (LLMs). In this thesis, we will bring further evidence to the debates about improving ToM and its presence in LLMs by comparing the decisions of LLMs and humans in a ToM task. We will also test if their decisions can be influenced, be made more alike, and if, overall, both the performance of humans and LLMs can be improved in a ToM task.

Language occupies a central spot when it comes to the debates concerning humans and ToM, especially when it comes to improving ToM abilities in humans. ToM performance correlates with language proficiency and specific language structures [17–20]. For example, [17] conducted a meta-analysis over 104 studies concerning language and ToM, and their results suggest language plays a role in ToM performance, as certain language abilities, such as command of sentential complements, explain the variance of ToM scores up to 44%. In like manner, previous studies suggest there is long and short-term improvement in ToM performance through training on sentential complements [19, 20]. Command of pronouns has also been suggested to predict ToM performance [21]. This suggestion is in line with previous theories according to which self-awareness, which can be quantified by pronoun use, could cause ToM [22, 23].

Contrastively, the main focus regarding ToM in LLMs is debating ToM's existence, both positive [1, 24] and negative evidence [9, 11] having been provided. What is not debated, however, are ToM's potential benefits in LLMs [25], such as improved collaboration between humans and robots [26–28], better performance in agents, self-driving cars, virtual tutors [9, 11, 29], better language understanding [29], and providing answers to many long-existing challenging AI problems [6], among others. Because of this, many initiatives tried improving [30] or building ToM models [31], making it a widely-researched topic [32] and a long-term goal [32]. For example, previous studies have explored various methods for ToM improvement, such as building specialized ToM networks [31], adding specific modules for ToM processing [33, 34], prompt eingineering [35–37], fine-tuning smaller models on answers from bigger LLMs that are better at ToM [38], and training models on ToM-related stimuli [34]. Thus, using linguistically inspired approaches to fine-tune models for improved ToM was not the main focus of previous studies [31, 33, 34].

When it comes to how LLMs and humans compare w.r.t. language use, there are several studies investigating how well humans and models align w.r.t. to different linguistic abilities [39–41]. For

example, LLMs were shown to follow similar tendencies w.r.t. reading behavior [39], grammaticality of sentences [40] or associations of concepts [41]. LLMs were also shown to pay attention or base their decisions on similar words and aspects as humans in text classification tasks [42–44], with models that replicate human behavior performing better [43, 43, 45].

Thus, there is a close link between language command, certain linguistic structures, and improvement in ToM as specific language structures have been shown to improve ToM in humans. However, despite such reported improvements, to our knowledge, previous studies did not test if participants could be even primed by exposure to such linguistic constructions to improve their performance on a ToM task immediately. In like manner, no previous study tried to improve ToM in LLMs by fine-tuning them on certain linguistic expressions. Even more, no previous study has compared the attention or decisions of humans and models in ToM, despite the improvements shown by such a comparison in understanding the model and its decisions better [46].

To address these shortcomings, we created an experiment that compares the decisions of humans and LLMs in a ToM task. In our experiment, we additionally test if exposure to linguistic constructions correlated with heightened ToM results in better performance for humans and models and closer decisions when comparing them. We will now intuitively introduce the task to give a general idea of what it intends. Upcoming chapters will explain the details of each part of the experiment, which amounts to three smaller experiments: the human experiment, the model one, and the experiment comparing humans to models.

1.1 Research idea

The research question of the current MA thesis is **Are decisions of LLMs and humans similar in a ToM task?** corresponding to the following hypothesis: **Decisions of humans and LLMs are similar in a ToM task**. The hypothesis is motivated by the fact the ranks of attention to words of models and humans were previously shown to be similar [42–44] when it comes to text classification. To compare humans and LLMs, we test both categories on a False-belief Task (FBT), and we look at how the importance paid to specific words changes if the models are fine-tuned or if the participants are primed with certain linguistic constructions.

In FBT participants make predictions about what are the beliefs or attitudes of characters in various contexts. FBTs test the ability to tell if the belief of one character holds for its reality (i.e. true belief) or if it is false (i.e. false belief), thus the task's name. In FBT each stimulus has a context, a question inquiring about the belief of one of the characters, and a set of possible answers. Thus, we create FBT stimuli to which humans and LLMs are exposed. However, to compare what is important for humans and LLMs, we need to measure their assigned importance to words in the FBT task.

Thus, for humans we use a self-paced reading task (SPR) where they read the stimuli at their own pace, word-by-word, with reading times recorded. We assume their reading times correspond to their decisions: the longer the reading time for a word, the more critical that world is for the answer given, as human reading behaviors correspond to the processing cost of read words [47]. Our human experiment also presents questions sometimes before the context of the stimulus, and not after, to exactly record what is important for participants in answering the question.

In comparison, models are exposed to questions and contexts as in a traditional extractive question and answering task (Q&A): given a context (the story of stimulus) and a question,

the model is asked to predict the correct answer. To weigh what is important for models, we use an explainability method called attribution maps. For each word in the context, we can calculate how important that word is for the models' decision to provide an answer. Afterward, we rank words in terms of highest attribution scores and reading times, thus mapping where each word stands in the decisions of models or humans. In this way, we do not compare reading times and attribution scores directly but rather compare how similar the ranks of humans and models are.

To test whether specific linguistic constructions improve performance or bring the ranks closer, we expose humans and LLMs to them. Humans are primed by the questions of the stimuli, which can be reformulated to contain a pronoun, sentential complement, or both. To 'replicate' the same priming effect in models, we fine-tune them on Q&A datasets where questions are modified to contain more pronouns and sentential complements. The models will also respond to the same questions as humans, which contain linguistic constructions in various degrees. Thus, we compare fine-tuned and non-fine-tuned models, as well as primed and control participants.

Thus, considering our main research question and the choices we made to operationalize it, ones we briefly described above, several research sub-questions appear:

1. Are FBT ranks of LLMs similar to human FBT ranks?
2. Are ranks of LLMs closer to human FBT ranks in fine-tuned models?
3. Does fine-tuning or priming improve FBT accuracy?

We will now proceed with introducing the contributions of the current thesis after which we will present how the thesis is organized.

1.2 Contributions of the current thesis

The contributions of the current thesis are the following:

1. Created an experimental design to test decisions in LLMs and humans. Our design is based on the assumption that human reading times and attribution scores can be compared. We introduce the Alignment Formula, a way to quantify and compare the decisions of LLMs and humans by their differences in assigned ranks to words.
2. Devised a method to fine-tune models for ToM implicitly.
3. Devised a method to modify implicitly existing datasets for ToM improvement using prompt engineering and by leveraging other LLMs for modifications.
4. Two ToM datasets of 8000 sentences containing two linguistic constructions correlated with better ToM performance in children: pronouns (i.e. *I, One*) and a sentential complement (i.e. *to think*).
5. A total of 12 fine-tuned models varying in the degree of exposure to the ToM-modified datasets.

1.3 Structure of the current thesis

The current thesis is organized as follows: Chapter 2 describes previous literature concerning the topics of the current thesis, aiming to make the reader familiar with the methods used in the current thesis at a theoretical level. Chapter 3 goes beyond theory and introduces to the reader how the tasks were practically implemented. Chapter 4 presents the data processing required for statistical tests, as well as their results. Chapter 5 discussed in depth the obtained results, while the conclusions, limitations, and future research suggestions are made in Chapter 6. Every chapter begins with a more in-depth summary of its subsections.

Chapter 2

Background

This chapter aims to lay grounds in the theoretical notions necessary to understand the following chapters, as well as offer a view of the overall previous research contributions that concern our topic. To achieve these two goals, Section 2.1 describes one of the most popular tasks used to test ToM, namely FBT, and Section 2.2 describes previous studies showing a link between language and ToM, with a discussion about sentential complements and pronouns. Section 2.3 reviews previous studies offering evidence for or against the presence of ToM in LLMs. Section 2.4 discusses how LLMs can be fine-tuned to align to human rationales, resulting in improved performance. Sections 2.5 and 2.6 describe methods to gather data about rationales of models or humans, namely attribution maps and the SPR task.

2.1 False Belief Task

Many tasks test ToM [13, 48], each having its potential downsides, e.g. not exclusively testing ToM [6, 49]. The tests target spontaneous or prompted ToM abilities [6], with the former naturally observing a participant and the latter making participants react to specific stimuli. While some argue prompted ToM observations are faulty due to them always being a reaction to a stimulus [6], it is essential to note that spontaneous tasks still observe reactions, but to more uncontrolled stimuli. A prevalent prompted ToM task is the FBT [4, 6, 24, 50, 51].

In FBT, participants predict the behavior of agents based on their true or false beliefs. The most typical scenario is when the location of an object known by one agent is changed when that agent is present or not [3]; see Table 2.1 for an example. Then, participants must predict where the agent will look for the object after its location changes [1]. The absence or presence of the agent at the change of location results in their false/true beliefs that (do not) correspond to reality. For example, in the aforementioned table, when Anne changes the location of Sally's book, the assumption Saly will have about the location of the book will be false.

As a task measuring ToM, FBT is highly debated, as people doubt if it should be used as a measure given its heightened difficulty for younger children [see 4, for a review and a meta-analysis of the performance of children on FBT by age]. The interaction with the researcher

[1] Note that several variations of the FBT exist. So far, we have discussed the *approach FBT*. Another variation is the *avoidant FBT* [3], where participants must predict a location an agent would want to avoid. Avoidant FBTs prove to be more difficult, while some argue that even normal FBT can be demanding for children.

Context	**Question**	**Correct Answer**	**Type of belief**
Once upon a time, there were two girls named Sally and Anne. Sally had a basket, and Anne had a box. One day, Sally put her book into her basket and left the room. While Sally was gone, Anne took the book out of Sally's basket and put it inside her box. After a while, Sally returned to the room and looked for her book.	Where do you think Sally will look for her book?	Her basket	False belief
Once upon a time, there were two girls named Sally and Anne. Sally had a basket, and Anne had a box. One day, Sally put her book into her basket. Anne took the book from Sally's basket and put it inside her box. After a while, Sally wanted the book.	Where do you think Sally will look for her book?	Anne's basket	True belief

Table 2.1: Typical scenario in the FBT task. The context is given so the participants can estimate the answers to the questions in the second column. The correct answers are provided under the *Correct answer* column, and the type of belief is given in Column *Type of belief*. False beliefs correspond to characters thinking their objects are in different places.

can also influence the response of participants in FBT experiments [6], making other methods, such as eye-tracking, more preferred. However, note that this downside is a consequence of any experiment in which subjects and researchers interact. Additionally, another possible downside of the task is that it asks about specific behaviors and has specific choices [6], indicating to the participant what is essential. However, FBT has been used several times in many research studies [14], offering much data for comparison. This is also why we are using it in our research, as it has been used often for both humans and LLMs. We will now proceed to discuss the link between language and ToM.

2.2 Relationship between ToM and Language

ToM is debated to be either an acquired or innate ability [14, 15]. Nonetheless, many studies show children acquire ToM in stages [14, 16], with older children scoring better in ToM tasks [4]. In this ongoing debate, the influence of language on the development of ToM [52] is strongly disputed [53]. For example, [53] describes several theories according to which language is linked to ToM because it is in itself a predictor of ToM abilities or because it might foster exposure to beliefs or attitudes through conversations and interactions with others. Similarly, [17] argue that language would be essential for ToM development as it is the primary tool by which children interact socially, listen to stories, and create narratives, aspects argued to lead to ToM development by [48, 54]. However, the leading opposing theory to this link is that because ToM tasks are mostly language-based [55], it would lead to the overestimation of the link between language and ToM [56]. This opposing argument is refuted by evidence obtained from testing other modalities for the FBT, i.e. visual, textual, or both, which do not result in different effects observed [55]. While this would show that not only language influences ToM, it also highlights the fact that the connection between them is not overestimated. We will now discuss the observations indicating that language affects ToM in more detail.

[17] conduct a meta-analysis on studies testing the link between language and false-belief understanding in typically developing children. The authors take all the correlations between language and false-belief understanding and find that, overall, their link is significant. Their study also suggests that in terms of the language ability that affects ToM more (e.g. semantics, syntax, memory w.r.t. sentential complements), syntax and remembering sentential complements are significantly correlated with explaining variances in FBT performance, with percentages of 29% and 44%, respectively. Note that sentential complements are verbs like 'to think' that are usually used to convey opinions of others, e.g. 'She thinks that (...)'. This would mean that the FBT scores across all participants could be explained by only looking at their ability to use sentential complements in 44% of the cases, which is a high percentage.

When we look at studies that concern the link between sentential complements and ToM, we generally observe that training on sentential complements results in implicit or *out-of-domain* improvement in ToM tasks. Note that out-of-domain improvement exactly marks improved performance obtained from training on stimuli not directly related to one task. For example, [18] test if the performance of pre-schoolers who failed ToM pretests increases if exposed to sentential complements, compared to ToM stimuli or relative clauses. The preschoolers, depending on the group, were trained for two weeks either on FBT stimuli and relative clauses or to describe scenes with sentential complements, receiving feedback. Their results show that the groups trained on the FBT stimuli and the sentential complements had significant and similar improved performance on FBT stimuli, as well as other ToM tasks for which they have not been trained for. Thus, such results suggest that training on sentential complements is as practical for improving ToM as training directly on ToM tasks. In like manner, [19] show that training on sentential complements improves long-term out-of-domain ToM performance for non-typical developing children. However, in their meta-analysis, [20] conclude that sentential complements are necessary only for developing ToM in atypical children, unlike typically developing children, where results would be more diverse. This remark aligns with the conclusion of [18], as sentential complements are beneficial but unnecessary to obtain improved ToM performance. However, note that in [20], 28 studies were selected for typical individuals and only seven for atypical ones. We can remark that the more considerable the number of reviewed studies, the bigger the chance to see more variation in the results they show.

Pronouns have also been shown to improve ToM. [21] investigate a phenomenon in language acquisition by which children allow the pronoun in example 1 to co-refer with the subject while not enabling the reflexive to co-refer to another person in example 2.

1. She likes her.
2. She likes herself.

The authors ask typically and non-typically developing children to match pictures to the sentences from the examples above, as well as to verbally describe images, which would eventually lead to them using a personal or reflexive pronoun. The children were also tested on FBT, working memory, and an inhibition task. In the FBT, children had to predict the beliefs of characters about objects as well as about other characters' beliefs, which are called second-order false beliefs [21]. The results of the study show that children who do worse in differentiating the non-reflexive nature of the personal pronoun in example 1 also do worse in the second-order FBT task, indicating a link between ToM and personal pronoun understanding. Note that this effect, according to the authors, evens out with age. Such results align with [22], according to which obtaining improved performance on examples such as 1 depends on ToM. If we consider a previously reported correlation between observing the self and observing others' mental states [see 51, for a review], with some arguing the self is used first to understand the others [23], the link between pronouns and ToM becomes less surprising. Even more, self-awareness can be quantified by the use of first-person personal pronouns [see 57, for a review], making it more plausible to assume a link between understanding the self and the use of first-person personal pronouns. We will now proceed in depth to describe positive and negative evidence for the presence of ToM in LLMs.

2.3 Large Language Models

2.3.1 LLMs and ToM

It is debated [1, 24] whether LLMs have [29] or do not have [9, 11] ToM abilities. We will go over the studies that show positive or negative evidence, and we will first review studies offering positive evidence.

Kosinski [29] argues ToM abilities could have arisen spontaneously in LLMs, as many other abilities have. Their argument is sustained by the fact that, when testing GPT-1, GPT-2, GPT-3, GPT-4, and BLOOM on new FBT stimuli, GPT-4 achieved 75% performance, a significant score. GPT-4 is also better than older and smaller models that did poorly. [58] also conclude that GPT-4 shows a high level of ToM as this model, ChatGPT, and `text-davinici-003` were exposed to several ToM stimuli. Their test dataset included new stimuli for the classic FBT scenario, as well as more varied scenarios in which the intentions or beliefs of humans need to be predicted. However, note that the authors acknowledge their study is not comprehensive, as they test a small number of stimuli, i.e. only one for each task is specified. Again, in the study of [12], GPT-4 achieves a significantly improved performance on ToM tasks. The model is tested along ChatGPT to estimate our bias of assigning a higher probability to information corresponding to our previous beliefs despite it being less likely. Not only did GPT-4 generally respond correctly to ToM tasks, but unlike ChatGPT, which always performed around chance, it also achieved in some subsets of the data 100%. Another study showing 100% accuracy for

LLMs is that of [13] where GPT-4, GPT 3.5, and LLaMA2 are tested on various ToM tasks, such as irony understanding or FBT, among others. On traditional FBT stimuli, all models and humans perform 100%. However, when FBT stimuli are modified to have transparent objects, for example, models decrease their performance, as shown by Ullman [9]. Note that the authors show humans follow a similar trend, arguing LLMs' decrease in performance could be due to their lack of understanding of certain concepts, such as being made of transparent material. On most tasks, GPT-4 outperforms humans. We will now describe the results of studies that showed negative evidence.

Many other studies argue ToM abilities are not present in LLMs. For example, unlike [29], [9] shows that GPT-4 performs much worse on FBT variations, a result also predicted by [10]. For example, the task used by [29] is a traditional version of the bag of Smarties. In this task, participants must predict what characters think the bag of Smarties contains. Participants' predictions should be based on their knowledge of what someone thinks typically is in the bag. When [9] adds variations to the task, such as making the bag transparent, the model fails almost by 100%. However, note that the study of [9] is limited in the number of models tested, as only GPT-4 was tested on the assumption that it has the best performance of SOTA models, thus reflecting the performance of the other models as well. More models should be tested to attest to such an effect across architectures, training data, or parameters. Similarly, [1] test LLMs on already existing ToM benchmarks, as well as a new one containing slightly modified FBT stimuli gathered from ToMI [59], and the papers of [29] and [9]. On ToMI or SocialIQA, models could perform around 70% with Chain-of-Thought prompting. On their comprised dataset, models do not exceed 30% performance. Because of this decrease in performance, their results suggest that models use spurious correlations to perform well on ToMI. Their results on slightly changed FBT are similar to those of [9], showing models are not robust to small stimuli changes. Out of the models tested, GPT-4 scores were the best.

Three variants of GPT-3 are asked to identify reactions in social interactions and beliefs in FBT using SOCIALIQA [60] and TOMI QA [59] in the study of [11]. Their results show that when identifying reactions in social interactions, the best models perform worse than humans with 20%, with bigger models not improving substantially to assume size can raise chances of human-like performance. In the FBT task, models achieve 60% accuracy at best. The authors report that models also start to be biased after their fourth time being prompted with the same stimulus by choosing the object last mentioned in the prompt. Another study is that of [15], where they test both humans and LLMs on FBT, showing GPT-3 differentiates false and true beliefs by assigning a lower probability to the true locations of objects in a false scenario. However, GPT-3 does reach only a 61% accuracy on the task, and it is not as sensitive as humans to the knowledge state of characters. Lower performance of models is also shown by [24], where they test if models such as Falcon, Pythia, GPT-2, and Llama assign different probabilities to true or false locations depending on the scenario. They show models assign, on average, bigger probabilities that characters will search objects in true locations, in true belief scenarios, close to 70% of the time. However, detecting false beliefs is more challenging for models, assigning a bigger likelihood to the false location only 50% of the time. Larger models were shown to statistically perform better than smaller ones, with the best model, Llama-30b, achieving almost 70% performance.

A study showing models to be outperformed by humans is that of [30], where models were tested to predict which character would benefit the most from finding the real location of an object if two or more characters look for it. While humans achieve 95% performance in more than 90% of the instances, models such as PaLM 2, ChatGPT 3.5, and GPT-4, achieve at best 50%. Models

are not only outperformed by adults but also by children. When comparing children of around 7 to 11 years old and models in the FBT, on a dataset containing first-order, second-order, and altered FBT stimuli, results of [8] show models perform better than children on first-order FBT stimuli, but do worse than them in second-order FBT stimuli.

With these studies revised, we see that the number of previous works pointing out a lack of ToM abilities is larger. Note also that these studies are generally obtained from more diverse methodologies, as they test more models in various tasks that are challenging. With this identified trend, we can wonder what can explain the lack of ToM abilities in LLMs. For example, Gurney et al. [6] argue that low results in ToM tasks in LLMs might be caused by only testing models by prompting, excluding spontaneous signs of ToM. As we have remarked previously, spontaneous ToM, though perceived as spontaneous, actually receives this categorization because the stimuli are not controlled for. Thus, even if we test the spontaneous signs of ToM in LLMs by a regular chat or other techniques, there is no such thing as genuinely spontaneous, as the model will always react to the input due to its architecture and how it is built to work. Even more, previous studies such as [24] test probabilities assigned to words, so models are not always tested only by prompting. [10] argue that the implicit manner in which ToM abilities appear in the training data of models is insufficient for acquiring them. The authors also say that the examples presenting explicit ToM scenarios are template-based, lacking discourse or syntactic complexity. This can result in a lack of generalization. The datasets available to fine-tune for improved ToM performance are also small, which can cause a lack of generalization in a model, as it will learn the dataset too well.

2.4 Fine-tuning after human values

In this subsection, we will discuss a particular trend in fine-tuning LLMs, i.e. fine-tuning after human values or rationales. As remarked in our introduction, no previous study aimed to align humans and models in ToM, which is the chosen methodology for the current thesis. The following studies aim to help the reader understand why alignment after human rationales can benefit models.

The topic of *alignment* between LLMs and humans has started to be more debated recently [61], especially w.r.t. ethical values or language processing. On the one hand, alignment could be helpful for models to have the same values as humans, such as safety and non-harmfulness [62], which is referred to as alignment with human values. This aspect becomes increasingly more important with the deployment of LLMs in real-life scenarios. On the other hand, the alignment of LLMs to humans and their processing of language has been shown to lead to better performance [43, 45] and to potentially lead to better user-computer interaction [46]. Still, it is not the only way improved performance can be achieved [63]. We will now discuss what previous works have shown regarding how much LLMs align with human tendencies.

Studies that test only the alignment of the output of LLMs and humans show they follow similar trends. For example, LLMs replicate human trends in judging the grammaticality of sentences in [40] as they find acceptable sentences more grammatical than unacceptable ones, similarly to humans. Similarly, [41] test if prompted LLMs map concepts from unrelated domains (such as *piano* to *animal*) in similar manners as humans. When comparing their explanations, humans and models report having based their decisions on aspects that are similar 83% of the time, with top LLMs being able to predict the top-rated associations from humans more in 25% of the cases. Note that testing the outputs of models and their alignment to humans is always less

reliable as the observed agreement might have been achieved because models were fine-tuned to meet human expectations in conversations or in outputs, thus possibly resulting in them giving responses that are expected by humans, which would mark an alignment with their preferences. Investigation into the inner workings of models is more reliable, which we will discuss next.

The internal representation of models can be compared to human rationales. For example, [39] test how well LLMs match the human brain activity on a reading task by comparing how closely the internal representations of models align with brain scans of people reading. The alignment between brain scans and LLMs is obtained by training a linear model from the features of LLMs to predict fMRI scans. The authors especially study the difference between non-fine-tuned models and instruct-tuned LLMs in their alignment. They show that instruct-tuning aligns models more with human brain activity, gradually improving alignment.

Another way to compare humans to the inner representation of models is to investigate what they pay attention to in various tasks. For example, [42] test how closely aligned the attention-maps of models are to the words humans pay attention to when answering a sentiment classification task. To obtain human judgments, humans were asked to highlight words that reflect the sentiment they decided a sentence has. The similarity between humans and models is defined as the similarity of the vectors that represent human and model choices, where the values in the vector can be either 0 or 1 (i.e. 1 representing an important word, while the vector and its length represent the words in the sentences to be classified). The results of the study suggest that the words identified by humans to be important were ranked as important in attention maps of models too. Humans and models also seem to pay attention to categories of lexical words similarly, with a bidirectional RNN replicating human attention scores best w.r.t. all aforementioned aspects. In a similar study, [44] compared the most important words LLMs, a logistic regression model, and non-expert humans based their decisions on in classifying narrative texts representing types of injuries. The data used for humans was eye-tracking data, a map showing important words for the decisions of models was used for open-resourced language models, and prompting was used for chat-based interfaces like Chat-GPT. Generally, in classifying the texts, the logistic model performed best, with ChatGPT varying depending on the labels to be predicted, from around 60% until almost 100%. Human reading times varied from 30% to 90%. The comparison was made between the top 5 words with the highest decision scores, the top 10 words prompted from ChatGPT as important for its decision, and the top 10 words with higher eye-tracking fixation times. Their results show that in all stories, the word that was first in ranks was common in the ranks of all models and humans. This trend declined for words lower in the rank. Thus, these studies show that models and humans generally pay attention to similar things in text classification or sentiment analysis.

However, [43] shows a different tendency. They investigated what humans pay attention to in a sentiment analysis task compared to models. The human data included eye-tracking data from an online experiment. Thus, the human attention map data would be determined based on what participants looked at. Their results revealed that when comparing human and model attention scores, models were observed to pay more attention to words related to sentiments, while humans would generally pay attention to more ambiguous words. They further study if the eye-movement features are present in the model behavior and find that the models replicating the human features closer also perform better.

After reviewing all these studies, we have to review what can bring LLMs and humans to align: direct and implicit (or out-of-domain) fine-tuning.

By training directly for alignment, we refer to fine-tuning models on human data to bring them closer to humans in certain aspects. For example, both [64] and [65] achieve human alignment by providing human rationales or human data to the models, and in the study of [66] or [67] attention maps of humans and models are closer if models have seen eye-tracking human data or attention maps of humans.

Contrastively, out-of-domain alignment follows to obtain an indirect improving effect by providing training data that is not directly linked to the task but can lead to improvements because it concerns aspects that were shown to be connected to the abilities required in the task. For example, [68] talks about the possibility that reasoning is improved in LLMs by training generally on language because language is considered to be linked to reasoning. In the current thesis, we test out-of-domain alignment as we chose to fine-tune models not directly on FBT stimuli but on linguistic structures that were correlated with improved ToM ability.

We will now describe the techniques used in obtaining the inner representations of models by describing the domain of explainability and attribution maps, an explainability method used in the current work.

2.5 Explainability

LLMs are often referred to as 'black-box' systems [69, 70] meaning that despite their good performance on a task, it is hard to tell what makes the models perform well, and if their performance is based on reasonable observations. Lack of transparency can lead to harmful behaviors of the model, especially in contexts where models are applied to real-life situations, as they can encourage harmful behaviors [71]. Explainability, as an area in artificial intelligence (AI), strives to bring clarity to the opaque nature of A.I. models. This goal aims to foster trust for users, as they better understand models, as well as to improve performance for models, as transparency indicates what their decisions are based on [69, 70, 72].

Methods used in explainability can be divided into two categories: local and global methods. Local explainability methods aim to show how models work w.r.t. a specific set of inputs. Because such a method shows only what features or words are essential for the model only when reacting to particular inputs [70], it offers a local explanation of how the model works [73]. Contrastively, global explainability methods aim to clarify what the models have generally learned or what specific components/layers know about inputs generally [70]. Global methods help us understand how models typically work. In contrast, local methods target more specific features and inputs [73], which is also why they will be used in the current work. With global methods, we would have to find an equivalent in the human experimental part of layers from models to humans, which is not possible given our current available design and resources. However, as [39] showed, a brain scan might prove a good equivalent for inner representations in models.

Local explanation methods based on features of models can be obtained by means of perturbation of the stimuli, using a gradient method, surrogate models, or decomposition [70, 74]. All of these methods are known as attribution methods [74] because they aim to assign a contribution, relevance score, known as an attribution value [75], for each feature (e.g. word) in the input. Often, these methods are visualized in so-called attribution maps [75]. First, we will explain what all the methods do and intuitively which is best for the current thesis. In the next section, we will offer more formal details about the technique used in the current work.

Perturbation-based methods Perturbation-based methods work on the assumption that if we take away or change a feature (e.g. word, token), we can see the importance the feature has in the model by observing the impact its absence has [74]. The worse the model does, the more important the feature is. The downside of this method comes from its assumption that features only have an individual effect, ignoring the fact that, more likely, every feature contains information about other ones since language is based on dependencies [74]. Another downside of this method is also its high complexity.

Gradient-based methods A less-computationally complex explainability method [74], the gradient-based method assumes that the contribution of one feature to the output is measured as the influence that feature has. Unlike in a perturbation-based method, gradients-based methods do not take away the feature but ask the model to perform a task (i.e. classify a text) and look inside the model (i.e. backpropagate) to see how features (e.g. words) are weighted in it. If the feature has a bigger weight, the model learned its contribution to be more important or have a bigger effect on that outcome. The coefficients (numbers) of the weights tell us how much they contribute and how the output changes w.r.t. that feature, making that feature a partial derivative of the model. Thus, this method is called the gradient-based method because when all features are present, we get a map of how the outcome changes w.r.t. all features, which together form the gradient [75]. There are several gradient-based methods, such as integrated gradients [72]. Gradient-based methods also have downsides, being still computationally demanding [74].

Surrogate models This method, as its name suggests, builds models as copies of other existing models. The 'surrogate' models should be equivalent to the original model and are used to test what is the decision region of a model [70, 74]. Some well-known surrogate model methods are LIME [76] and SHAP [77], according to [70].

Decomposition methods This feature-based method assigns a relevance score for input by breaking down its relevance in the last layer or across layers in the network [70, 74]. A well-known method from this category is Layer-wise Relevance Propagation [78].

Our task, as mentioned in the introduction, is to compare what contributes more to the decisions of humans and models in the FBT. Because of this purpose, only the assumption of gradient-based methods is suitable for our task: the human experiment will test how the words contribute to humans' decisions by reading times, which would be a direct equivalent of attribution scores. The other methods are not suitable because to build an equivalent for perturbation-based methods, we would have to conduct several human experiments in which we would take words out of the stimuli, one at a time. Regarding surrogate methods, we cannot assume LLMs are surrogate models for human reading times or their decisions. That is especially because our task has not been tested before, and we do not know how well models replicate human reading times in an SPR version of the FBT. In contrast, for the decompositional techniques, we do not have an experimental equivalent of relevance for humans.

LLMs are more complicated to explain given their larger size and their better, more extended performance to a variety of tasks [74, 79]. Knowing this, we realized that not all the available implement gradient-based methods would work for LLMs. Thus, to choose which type of gradient to calculate, we first reviewed the libraries available for LLMs; more explanations about our review can be read in the methods section 3.2.2. Considering our task (i.e. Q&A) and the available libraries, only one option remained: integrated gradients [72].

Integrated Gradients (IG) is a type of gradient-based method that differs slightly from the intuitive explanation we gave when we introduced general gradient-based methods. In the given

intuitive explanation, we mentioned that the coefficients, i.e. the weights, are considered to be the contributions of features, i.e. gradients. IG calculates these coefficients slightly differently.

Taking a baseline, e.g. a 0 embedding vector, and every step taken from it to get to an input and how its features are represented, IG is calculated as an average of all contributions made by each step taken towards the input features. For example, consider a feature vector that has only 2 values of 1 in it, with all other values being zero. In this case, compared to the baseline vector, the difference between the two is those only two values. IG will first change one of the different values from 0 to 1. Then, the gradient of the vector will be calculated after this small change in value. Afterward, it will also change the second different feature to 1, this vector actually being the vector of all input features. A gradient will be calculated for it. However, the gradient of the input feature will not be that after the second change of values but a sum between the two vectors: the one after the first and the one after the second change (which actually represents the final input vector). This is a simplified case of a vector of input features, as the vector might have to make many different slight changes to get from the baseline vector to the input one. Thus, IG is the averaged contribution of all steps taken from the baseline to the input, where the contribution is formally known as the gradient (how a slight change in the input affects the output for all steps taken from the baseline). This explanation is based on the paper of [72], where formulas of IG can also be reviewed.

As a specific type of gradient-based attribution method, IG has many benefits: it applies to various models while also being consistent across various implementations [75]. IG is a method that has a property of Sensitivity [72, 75], which is a form of sanity check: the difference between the output obtained from the input and that from the baseline can be reconstructed by summing up all the IG of the features from the input. This intuitively makes sense because the output should be a consequence of all the contributions of IG as a reaction to the input. This type of sanity check does not hold for other gradient-based methods. However, IG is also computationally costly, given the amount of gradients calculated. Thus, for bigger models, it takes longer to compute.

We will next discuss the equivalent for IG in the experimental design for human decisions, that is, human reading times obtained from SPR.

2.6 Self-paced reading task

In SPR, participants read sentences on a computer screen [80] at their own pace, making sentences appear word-by-word, having their reading times being recorded [80, 81]. To facilitate reading at their own pace, the participants press a button to reveal the sentence [81], see an example in Image 3.4. Reading times of words reflect their processing time, with longer times signaling more difficult processing [80] or more engagement [56].

There are several types of sentence segmentations for SPR, with word-by-word segmentation offering fine-grained information and excluding potential biases due to specific group segmentation [81]. In contrast, group segmentation has more ecological validity [81]. The number of segments and their composition, e.g. word 1 is a noun, are recommended to be consistent across stimuli [81]. Depending on the type of SPR, pressing the button for a new segment makes the old segment be kept (*cumulative* SPR) or disappear (*noncumulative* SPR). The positioning of the segments on the screen also differs, with segments appearing in the center of the screen (*centered* SPR) or linearly (*linear* SPR) like in a sentence [81]. Usually, SPR is used to test ambiguities,

language violations, or dependency relationships at a distance [81]. Across stimuli, segments are controlled for length, syllables, and frequency [81] to prevent longer reading times due to uncommon words or differences in lengths of words [82]. In SPR, to avoid participants focusing on the critical sentences, [83] distractors come after each stimulus, being similar to them [81]. Stimuli are also followed by balanced comprehension questions to ensure the engagement of the participants [81].

There are some disadvantages associated with the SPR task. For example, SPR seems unnatural when compared to normal reading [84]. Cumulative SPR allows participants to reveal first the whole sentence and then read it [82, 85], while centered SPR is unnatural due to its positioning [81]. Additionally, in non-cumulative SPR, readers cannot look back, while the cumulative version does not offer a way to record their look-backs [86]. However, SPR also has many advantages. One of them is that SPR allows participants to move at their own pace [80]. Additionally, it is easy to implement and also widely used [81, 84], thereby offering many results for comparison, the main reasons for which it was chosen for the current thesis. In the current thesis, we use noncumulative centered SPR to test decisions in FBTs, which has not been previously investigated.

The current chapter reviewed the following main topics: the FBT task, the relationship between language and ToM found in humans, ToM abilities of LLMs, ways of testing and fine-tuning models for human alignment, as well as the two main techniques that will be used to obtain data about decisions of people and LLMs, namely attribution maps and SPR. The next Chapter describes how the tasks were implemented.

Chapter 3

Methodology

This chapter describes in-depth the methods used in the current study. The chapter is organized as follows: Section 3.1 describes the creation of the new FBT dataset by synthetic data generation, the division of the stimuli across groups of participants, and how questions were rephrased with specific linguistic constructions. Section 3.1.2 describes how the SPR experiment was implemented, while Section 3.2 how fine-tuning was achieved w.r.t. model choice in 3.2.1.1, modifications brought to already existing datasets in 3.2.1.2, hyperparameters for fine-tuning in 3.2.1.3, models' evaluation in 3.2.1.4, and obtaining attribution map scores in 3.2.2.

Note that even though we explain the creation of the FBT stimuli as part of the human experiment, the same stimuli were used to evaluate models. Explaining the creation of the FBT stimuli in the human experiment part was a choice made based on the fact that the FBT stimuli were modified, especially considering aspects tied to the SPR experiment, such as making the FBT stimuli shorter not to make the SPR task too difficult.

3.1 Human Experiment

3.1.1 Creating the FBT Stimuli

We created a new dataset for the FBT using synthetic data generation to avoid previous exposure to stimuli that results in increased performance [87, 88]. Consequently, we used GPT-4 to create similar stories to those of [1], given their increased difficulty, unlike benchmarks such as SOCIALIQA or TOMI. The synthetic data generation presupposed taking the dataset from the study above and prompting the model with the prompt in Figure 3.1.

The generated stories were similar to those of [1] in style, but they did not use the same stories, places, or people, making the generated dataset original. However, many generated stories had to be manually checked and modified to avoid models using their lack of plausibility as a cue for which belief is correct. For example, one of the stories assumed that items that would look the same, such as honey and caramel, could be differentiated by their look. In some generated stories, agents could even see through objects specified to be non-transparent. The generated stories were also modified to all contain only 35 words so that the reading times could be compared after the SPR task. To illustrate the resulting stimuli used in the SPR task, and later used also in the FBT language model task, see Table 3.1 below for an example. Each stimulus contains three

```
Instruction: I will give you a story. Create a similar story.
Input: "A bottle filled with perfume sits on the counter of a department store.
    There is no water inside it, but one of the employees labelled it "water" instead of
    "perfume" as a joke. The employee walks up to the  counter and notices the bottle.
    He reads the label. He opens the bottle and looks inside. He can clearly see that it
    is full of _"
Output:
```

Figure 3.1: Prompt used to generate similar stories to those of [1]. The input example is taken from the dataset provided by the paper.

parts: the story/context, the question, and the response variants. From now on, when we want to refer only to the story, we will use the term *context*. When we want to refer to the whole stimulus, including its question, we will use the term *stimulus*.

The story	In a grocery store, there is a non-transparent jar filled with honey. An employee wrongly labeled it 'wheat' instead of honey. A customer reads its label, opens the jar, looking inside. The customer closes it.
The question	What does the customer think is in the jar?
Variants of response	honey; wheat

Table 3.1: An example of three parts that form a stimuli in our FBT.

Each context was altered to have a false and true belief version. One example of the alternation is given in Table 3.2.

Context	**Belief**	**Question**	**Answer**
In a grocery store, there is a non-transparent jar filled with honey. An employee wrongly labeled it 'wheat' instead of honey. A customer reads its label, opens the jar, looking inside. The customer closes it.	True	In the end, what does the customer think is in the jar?	honey
In a grocery store, there is a non-transparent jar filled with honey. An employee wrongly labeled it as 'wheat' instead of honey. A customer sees the jar, reads its label, and decides to buy it.	False	In the end, what does the customer think is in the jar?	wheat

Table 3.2: Examples of how the contexts vary considering their expressed belief. What makes contexts have a true or false belief depends on the question asked, highlighted in the column *Question*.

Besides the stimuli, we created fillers by prompting GPT-4 to generate stories about characters' wishes (e.g. what someone would wish to be in a box). For fillers, we also considered using sentences or scenarios concerning religion by asking participants to estimate what an agent

thinks based on such beliefs. In this way, we could have covered very elegantly the true research question of the study by the ambiguity between a religious belief and a belief about an object's location. However, we eventually ruled out this possibility, given religion can be a sensitive topic for many participants. Nonetheless, these stimuli will also be made available and can be used further in studies that deal with religion, language, and beliefs. Unlike the stimuli, the context in fillers did not alternate between true and false beliefs; their purpose was only to distract participants from the study's research question.

Previous studies have shown that certain linguistic constructions, such as command of pronouns or exposure to sentential complements (e.g., believe, think), improve performance in ToM tasks [89, 90], see Section 2.2 for more details. As no previous study tested if improvements in ToM tasks can be obtained immediately, we decided to prime participants by having three different types of questions containing the aforementioned linguistic constructions in various degrees, namely the Control question, the Agent question, and the Reader question, see Table 3.3 for examples of them. A Control question does not contain any specific linguistic structure shown to improve ToM and only inquires about what could be in the object that is the main focus in the context, e.g., in the jar or box. For instance, the Control question in Table 3.3 only inquires about the object, i.e. the jar. The same table shows that the Agent question is obtained by rephrasing the Control question to contain the sentential complement 'think'. In contrast, the Reader question is obtained by adding the first personal pronoun 'I', besides the sentential complement 'think'. Note that the Reader question was phrased in a lengthy manner to sound natural to participants. From now on, we will refer to questions only by the Control Reader or Agent question.

The Type of Question	**Example of the Type of Question**	**Linguistics Construction Correlated with heightened ToM**
The Control Question	What is in the jar?	None
The Agent Question	In the end, what does the customer think is in the jar?	Sentential Complement
The Reader Question	Ask yourself: What do I, the reader, think is in the jar, according to the customer, in the end?	Pronoun, Sentential Complement

Table 3.3: Examples of three different types of questions and their associated linguistic constructions. Note that the Control question is not formed to have any specific structure.

Depending on its type, the question comes before or after the context. Because we want to test how specific linguistic constructions affect the correctness of responses and reading times, the Reader and Agent questions always come before the context. In this way, participants will know beforehand what to follow and expect and what to pay attention to. This might make them pause more for the words in the context that are important to answer the question. Contrastively, the Control question comes after the context and is meant to show what the reading times of the participants look like when reading the sentence without a specific question in mind.

The response variants were binary and were either the correct answer or the other plausible but wrong possibility. For example, the response variants for the contexts and questions in Table 3.1 are honey and wheat. The full stimuli and fillers are in the Appendix A.2.

The text generation to form the stimuli resulted in 18 contexts, all with false and true belief versions and 6 fillers. We first decided to opt for a Latin-square design to divide the stimuli across groups of participants. This would mean that each participant sees each context only with one type of question and one belief. Thus, note that when we actually refer to a stimulus, we refer to a combination of types of questions and beliefs of the context, e.g. participant one sees stimulus 1, which is the first context with the question Agent, in its true belief version, and participant two sees stimuli 1, which is the first context, with question Reader, in it false belief version, and so on. Note that all the stimuli we referred to so far are actually combinations of types of questions and beliefs of the context.

However, because the contexts had been proven to be difficult even for humans [60], the SPR experiment would be challenging to complete if we exposed each participant to all stimuli. Consequently, we decided to split the stimuli into three subgroups, each of 6 stimuli, and we would have participants see one of the three subgroups. The participants would see the same 6 fillers in all subgroups, resulting in 12 items per test per participant. Given we had three types of questions and also true or false beliefs, we had to make sure each participant saw an equal number of stimuli containing true and false beliefs, an equal number of Agent, Reader, and Control questions, especially to have enough data for each observation. This resulted in the following division of stimuli across groups:

Table 3.4: Table for Group 1

Subgroup 1	**Subgroup 2**	**Subgroup 3**
7 persons	7 persons	7 persons
context 1a AgentQ	context 1b ReaderQ	context 1a ControlQ
context 2b ReaderQ	context 2a ControlQ	context 2b AgentQ
context 3a ControlQ	context 3b AgentQ	context 3a ReaderQ
context 4b AgentQ	context 4a ReaderQ	context 4b ControlQ
context 5a ReaderQ	context 5b ControlQ	context 5a AgentQ
context 6b ControlQ	context 6a AgentQ	context 6b ReaderQ

Table 3.5: Table for Group 2

Subgroup 4	**Subgroup 5**	**Subgroup 6**
7 persons	7 persons	7 persons
context 7a AgentQ	context 7b ReaderQ	context 7a ControlQ
context 8b ReaderQ	context 8a ControlQ	context 8b AgentQ
context 9a ControlQ	context 9b AgentQ	context 9a ReaderQ
context 10b AgentQ	context 10a ReaderQ	context 10b ControlQ
context 11a ReaderQ	context 11b ControlQ	context 11a AgentQ
context 12b ControlQ	context 12a AgentQ	context 12b ReaderQ

As it could be seen, each context would be seen by 21 participants, while each subgroup of participants would see two stimuli with AgentQ, two with ReaderQ, and two with ControlQ. Three contexts had false beliefs, while the other three had true beliefs. Note that because we have three types of questions and only two types of beliefs, we would not be able to get both beliefs seen by an equal number of participants for each context. For example, context 1 would have 14 participants seeing its true belief while having only 7 for the false one. However, across groups, we would have, in the end, an equal number of true and false beliefs seen by participants. Most importantly, we will have Control questions equally divided between false and true beliefs across groups. For example, in Group 1 the Control questions will be grouped with three true

Table 3.6: Table for Group 3

Subgroup 7	**Subgroup 8**	**Subgroup 9**
7 persons	7 persons	7 persons
context 13a AgentQ	context 13b ReaderQ	context 13a ControlQ
context 14b ReaderQ	context 14a ControlQ	context 14b AgentQ
context 15a ControlQ	context 15b AgentQ	context 15a ReaderQ
context 16b AgentQ	context 16a ReaderQ	context 16b ControlQ
context 17a ReaderQ	context 17b ControlQ	context 17a AgentQ

Table 3.7: **Tables for Group 1, 2 and 3** *Questions*: *AgentQ* refers to the Agent question, *ReaderQ* is the Reader question, and *ControlQ* is the Control question. *Small letters* The small letters indicate the belief of each context, with 'a' for *true* belief and 'b' for *false* belief. The same meaning for letters and questions holds for all nine subgroups.

and three false belief contexts. This will facilitate an equal comparison between contexts with true and false beliefs, as we take the Control question as the baseline. When we look at fillers, as in Table 3.8, we observe that they were distributed similarly. Because fillers do not have two beliefs, they do not differ w.r.t. their context across groups but only w.r.t. the questions they were presented with. Note that the table also shows which subgroups were assigned which contexts of fillers and questions. Note again that a filler is formed as a combination of a question, a context, and variants of responses (which are not part of the table). To avoid confusion, the context of fillers is represented differently from that of the stimuli, e.g. *contextF*. If we refer to context, we will refer to the one coming from a stimulus, and if contextF is used, it refers to one coming from a filler.

Subgroups 1, 4, 7	Subgroups 2, 5, 8	Subgroups 3, 6, 9
21 persons	21 persons	21 persons
contextF 19 AgentQ	contextF 19 ReaderQ	contextF 19 ControlQ
contextF 20 ReaderQ	contextF 20 ControlQ	contextF 20 AgentQ
contextF 21 ControlQ	contextF 21 AgentQ	contextF 21 ReaderQ
contextF 22 AgentQ	contextF 22 ReaderQ	contextF 22 ControlQ
contextF 23 ReaderQ	contextF 23 ControlQ	contextF 23 AgentQ
contextF 24 ControlQ	contextF 24 AgentQ	contextF 24 ReaderQ

Table 3.8: Subdivision for fillers per subgroups of stimuli. Each column represents the fillers assigned to certain subgroups, as well as the number of persons that were supposed to be exposed to them. Note that *AgentQ* refers to the Agent question, *ReaderQ* is the Reader question, and *ControlQ* is the Control question. *contextF* represents the context/story of the Filler. When we say a subgroup was exposed to 6 fillers, we refer to the fact that it was exposed to a combination of contextsF and questions, as well as variants of responses, parts that form a Filler.

3.1.2 The SPR Experiment

The SPR experiment was reviewed and approved by the Faculty Ethics Assessment Committee of the Faculty of Humanities of UU. The SPR experiment was set up and held on the `PCIbex.farm` platform [91] only after receiving the ethical approval of the committee. The SPR experiment used the stimuli and fillers described in Section 3.1.1 and was anticipated to last a maximum of 12 minutes. The data collected included the choices and reading times of the participants.

Because no personal or demographic data was collected, the SPR task was anonymous. Each participant was compensated a total of 1.3 GBP for the completion of the experiment.

The experiment starts with the letter of consent in which the participants are informed about the nature of the task, the time it will take to complete it, the compensation received for it, the right to withdraw at any time, the anonymity of the collected data, its protection and storage and the conditions on which the data can be shared with others. The experiment cannot proceed without the participants' consent; see the letter of consent in Figure A.1 in Appendix A.

The next page gives instructions to the participants. They are going to 'read texts and answer questions about them', questions which could appear before or after the texts, what we refer to as contexts. The words in the texts are also going to appear word-by-word. Because of this, the participants are advised to read carefully and not to rush. Check Figure A.2 to read the instructions received by the participants.

Next, the participants get a practice session where more instructions, including examples of the stimuli. For example, a question appears in Figure 3.2, while the context appears in Figure 3.3). The variants of responses appear in Figure 3.4. The aforementioned images also showcase the steps one participant has to take when a question comes before the context. The experiment starts after the practice session.

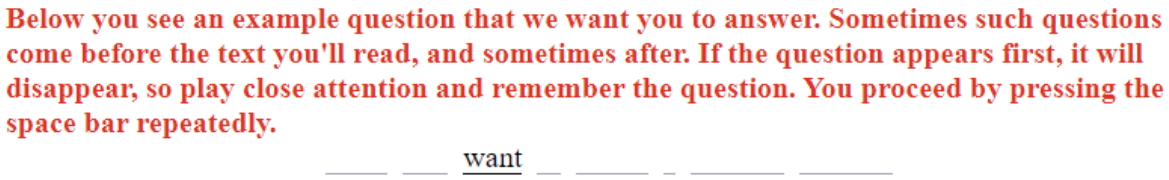


Figure 3.2: A question when it appears before the context, example part of the practice session

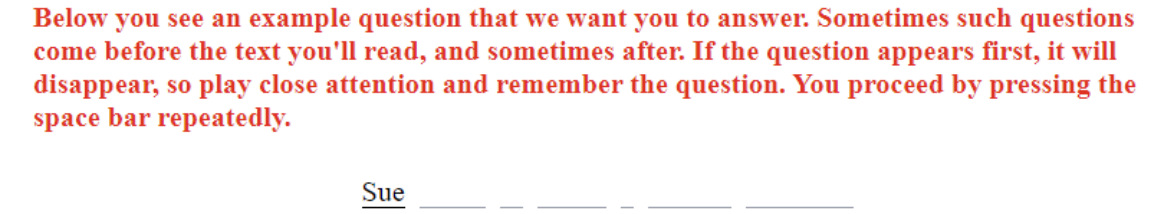


Figure 3.3: The context when a question appears before it, an example from the practice session.

After one participant completes the experiment, the randomization should assign the next new participant to the next subgroup. We piloted the experiment with two native English speakers. The pilot version of the experiment contained variants of the contexts of 55 words each. The main feedback received was especially regarding the difficulty of the task, which resulted from

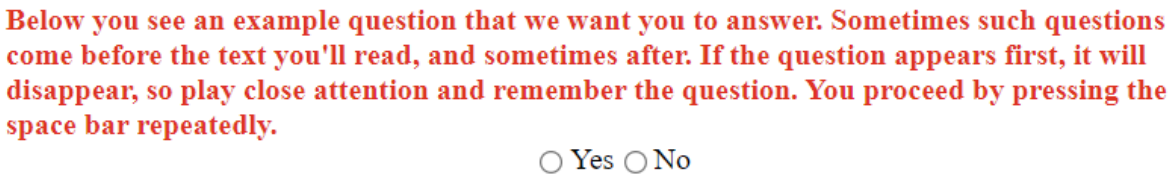


Figure 3.4: Variants of response when a question appears before the context, an example from the practice session.

the combination of disappearing words and lengthy stimuli. After the input from the pilot study, we simplified the contexts to 35 words. These stimuli were eventually used in the experiment.

Considering recruitment, the participants were recruited on Prolific. They had to be native English speakers and over 18 years old. The experiment took approximately one day to complete, and 82 participants participated.

3.2 LLMs

3.2.1 Fine-tuning

3.2.1.1 Models

Because the FBT is essentially an extractive Q&A task, i.e., given a context, the model has to extract an answer to a question, we decided to fine-tune already existing Q&A systems. We chose the models we would fine-tune based on three criteria: resources, popularity, and compatibility with the library `transformer-interpret`, used for obtaining attribution maps.
Considering our computing resources (i.e., Google Colab Pro), big LLMs such as Llama 8b or 13b [92] could not be fine-tuned due to lack of RAM. Out of all the other smaller available models, many of the models that could be used were further excluded as the library used for obtaining attribution maps was not compatible with them; see Section 3.2.2 for a detailed explanation of the choice of the library and the notable problems of the currently available libraries for attribution maps. Thus, even if the smallest version of Llamma 7b could be fine-tuned, it would not be compatible with the `transformer-interpret` library. Afterward, we used HuggingFace to search for the most Q&A downloaded models. Most of the top downloaded models were variants of RoBERTa [93] and BERT [94] fine-tuned on the SQUAD2 dataset, see 3.2.1.2 for a description of the dataset. Thus, we chose for further fine-tuning `deepset/roberta-base-squad2` and `deepset/bert-large-uncased-whole-word-masking-squad2`.

RoBERTa and BERT are transformer-based models, using attention to assign weights to important input parts [95]. These models, when trained, are asked to predict randomly masked words from the input. While BERT requires more computational power, RoBERTa requires less; see [95] for more formal details about these models' architectures.

3.2.1.2 Dataset Creation for Fine-Tuning

For fine-tuning the models further, the SQUAD2 dataset [96] was chosen. By choosing SQUAD2, which is the dataset on which our models were fine-tuned, we could isolate the effect of slight modifications brought by specific linguistic constructions, as the models are familiar with the style of this dataset. The SQUAD2 dataset contains 150.000 answerable and unanswerable questions, where an *unanswerable* question cannot be answered solely by extracting information from the context. For each question in the SQUAD2 dataset, a context is given to extract its answer, as well as which is the correct answer (represented by the *text* key in the 'answers' entry), where the answer starts, and if the question has one. For example, a typical entry in the SQAUD2 dataset looks like:

```
{
    "answers": {
        "answer_start": [94, 87, 94, 94],
        "text": ["10th and 11th centuries", "in the 10th and 11th
   centuries", "10th and 11th centuries", "10th and 11th centuries
   "]
    },
    "context": "The Normans (Norman: Nourmands; French: Normands;
   Latin: Normanni) were the people who in the 10th and 11th
   centuries gave their(...)",
    "id": "56ddde6b9a695914005b9629",
    "question": "When were the Normans in Normandy?",
    "title": "Normans"
}
```

The dataset has two available sub-datasets, i.e. data for training (130319 questions) and validation (11873 questions). From each of these sub-datasets, we sampled 8000 questions, each group having 4000 answerable and 4000 non-answerable questions. We sub-sampled only 8000 questions from both sub-groups because the validation sub-group only allowed this number, given its smaller length. By sub-sampling, we mean that the data corresponding to only 8000 questions (e.g. the answers, context, id, question, and title) was kept. Thus, we formed two data sub-samples, henceforth referred to as train-8k and dev-8k. These two datasets were used for fine-tuning, so a model would be fine-tuned on data from a dataset that it has been previously trained on and one that it has not. Note that the choice of sub-sampling from both the validation and the training parts of the SQUAD2 dataset was made to test if fine-tuning on the data coming from the train dataset, which models have seen before, affects their performance compared to the one coming from the validation data. However, the models are only trained once on original non-modified sub-sampled data from the train or validation datasets, to have them as a comparison for when we will train the models on datasets in which the questions contain pronouns or sentential complements, as we will explain the next paragraph. By fine-tuning on original non-modified sub-sampled data, we manage to differentiate by comparison the effect data with linguistic constructions has on models. Lastly, from the validation sub-dataset, we have also sampled an evaluation sub-sample of 2000 questions, equally divided between answerable and non-answerable ones, henceforth named eval-2k. See Table 3.9 for an overview of the aforementioned sub-sampled datasets.

To test if certain linguistic constructions improve performance in LLMs, pronouns and sentential complements were added to the train-8k and dev-8k.

We remind the reader that in the SPR experiment, participants would be primed by modification of the questions with pronouns and sentential complements. To achieve a similar priming effect, we modify the questions of the train-8k and dev-8k with pronouns and sentential complements into two possible types of questions: *I-think* questions and *One-think* questions. For example, the question from the original SQUAD2 dataset, which is the equivalent of the Control question, illustrated in the 'Original question' section in Table 3.9, is modified in the I-think question or the One-think question by adding either the pronoun 'I' or 'One' and the sentential complement 'think'. The One-think type of question is meant to be the correspondent of the Agent question from the SPR experiment part of this thesis, as the I-think question is for the Reader one. Slight modifications appear between the human experimental questions and those for LLMs as the 'Ask yourself' or 'the reader' would be unnatural in a model only trained on Q&A data with no specific instructions or training data specifying that the model is the reader. In the questions for fine-tuning models, the agent from the Agent question is replaced by 'one', as it would always be tricky to find an agent from some factual questions like the original question in Table 3.9, where an agent to emit an opinion or belief about a fact might not be part of the context.

Original question	When were the Normans in Normandy?
I+think question	When do I think the Normans were in Normandy?
One+think question	When does one think the Normans were in Normandy?

Table 3.9: Examples of different questions for fine-tuning.

GPT-4 was used to modify each original question in the I+think and One+think types of questions. The API of the model was used with the prompt in Figure 3.5, where 'I think' in the output differed considering what the output should contain (i.e. the pronoun 'one' or 'I'). Each new question to be modified would replace 'In what city and state did Beyonce grow up?' in the prompt to create an iterative generation process.

```
messages_prompt = [
    {
        "role": "system",
        "content": "introduce personal pronouns",
    },
    {"role": "user", "content": 'input: When did Beyonce start
   becoming popular? output: When do I think Beyonce started
   becoming popular? input: What areas did Beyonce compete in when
   she was growing up? output: What do I think were the areas
   Beyonce competed in when she was growing up? input: In what city
    and state did Beyonce grow up? Output:'
    }
]
```

Figure 3.5: The prompt used to modify original SQUAD2 questions with GPT-4.

After the questions were modified, they were replaced in train-8k and dev-8k, resulting in two new datasets for fine-tuning. Table 3.10 classifies the datasets obtained after data generation, as well as the original subsamples (train-8k and dev-8k) that were also used for fine-tuning to

isolate the changes created by the linguistic constructions, from that of the fine-tuning on data itself. The table also mentions the eval-2k dataset.

Name dataset	Type	Origin	Pronoun	Sentential Complement	Number of Questions	Possible Answers	Unpossible Answers
train-I	modified	SQUAD2-train	I	think	8000	4000	4000
train-One	modified	SQUAD2-train	One	think	8000	4000	4000
dev-I	modified	SQUAD2-dev	I	think	8000	4000	4000
dev-One	modified	SQUAD2-dev	One	think	8000	4000	4000
train-8k	unmodified	SQUAD2-train	none	none	8000	4000	4000
dev-8k	unmodified	SQUAD2-dev	none	none	8000	4000	4000
eval-2k	unmodified	SQUAD2-dev	none	none	2000	1000	1000

Table 3.10: Datasets Classification. The columns represent the following aspects: *Name dataset* represents the name of the dataset; *Type* represents if the dataset contains unmodified data (part of the original SQUAD2 dataset) or not; *Origin* represents the sub-dataset the data comes from; *Pronoun* represents what pronouns are part of the dataset, if any; Columns *Number of Questions*, *Possible Answers*, and *Impossible Answers* represent the numbers of questions that are part of the datasets, as well as how many of those questions are answerable based on the context provided, and how many are not.

3.2.1.3 Fine-tuning

`deepset/roberta-base-squad2` and `deepset/bert-large-uncased-whole-word-masking-squad2`, henceforth called the baseline models, were fine-tuned each on all the datasets in 3.10, except on eval-2k, using the Q&A pipeline from HuggingFace, resulting in 6 fine-tuned models for each baseline model. All fine-tuning datasets were formatted as Dataset objects and then tokenized with word masks, given both models require word masking for fine-tuning. The hyperparameters for fine-tuning for all models are shown in Figure 3.6 and were adapted to the specified hyperparameters of `deepset/roberta-base-squad2` from HugginFace to exclude any differences due to training on different parameters. The only difference was in the training batch, which was adapted to the computing environment. Note that `deepset/bert-large-uncased-whole-word-masking-squad2` did not have its hyperparameters specified in the model card on HuggingFace, so the hyperparameters of RoBERTa were also used for the BERT model.

```
train_args = TrainingArguments(
    output_dir="mdzrg/name_of_the_model",
    learning_rate=3e-5,
    per_device_train_batch_size=4,
    per_device_eval_batch_size=4,
    num_train_epochs=2,
    weight_decay=0.01,
    push_to_hub=True,
)
```

Figure 3.6: Training arguments for the fine-tuned models adapted after the hyperparameters specified for the `RoBERTa` model on HuggingFace.

As it can be seen from the 'True' value of the `push_to_hub` argument in Figure 3.6, all models were uploaded on HuggingFace to avoid a shortage of memory and to facilitate the evaluation process. These models can be accessed and tested on other stimuli.

3.2.1.4 Evaluation

Firstly, we replicated the evaluation results of the original models specified on HuggingFace, i.e. `deepset/roberta-base-squad2` and `deepset/bert-large-uncased-whole-word-masking-squad2`, on the full validation SQUAD2 dataset as a form of reassurance that our future results will be correct. However, when using the HuggingFace interface, we could not obtain the same evaluation scores as those specified for model cards on HuggingFace. The difference in results might have been caused by how HuggingFace loads particular models for evaluation on Q&A answering. Thus, we used the `FARM.reader` to load models, which resulted in slight differences in replicated evaluation scores as it has a word-based tokenizer, which is different from the HuggingFace interface that loads models with a token-based tokenizer. The replicated results, as well as the initial reported values for each model, can be seen in Table 3.11.

Model	**Reported results from HuggingFace**		**Replicated results**	
	EM	**F1**	**EM**	**F1**
roberta-base-squad2	79.87	82.91	82.22	84.30
bert-large-uncased-whole-word-masking-squad2	80.88	83.87	83.06	85.44

Table 3.11: Replicated scores of baseline models evaluated on the full dev-SQUAD2 dataset. *Reported results from HuggingFace*: the evaluation scores reported in the model cards on HuggingFace for each model; *Replicated results*: results obtained for replication using the FARM.reader. *EM* Exact match of the predicted answer with the gold standard in the dataset. Note that the scores are slightly different (i.e. higher) due to the difference in tokenization.

After that, all models were evaluated on two datasets: the eval-2k dataset and the FBT dataset. Note that in the FBT dataset, the questions for the stimuli are the same as those for the human experimental task to facilitate the comparison of attribution scores and reading times. Thus, models are evaluated on the Reader questions that contain: 'Ask yourself: (...)', despite not being trained on questions like that. The evaluation results are shown and discussed in Section 4.2 of the next chapter. To ensure our results are correct, we performed a sanity check: we took the rest of the dev dataset not included in eval-2k and tested the baseline models on it. Afterward, we summed the scores of the models on the two dev sub-samples (eval-2k and the rest of the dev dataset), and we divided it by 2, obtaining the replicated results in Table 3.11.

3.2.2 Attribution Maps

After evaluation, the attribution maps of the models were processed using the library `transformers-interpret` w.r.t. to the FBT stimuli. There are many available libraries for calculating gradient-based attribution maps. However, if we look at Table 3.12, we observe they are poorly documented or do not support well evaluating fine-tuned models from Huggingface.

Our chosen library is specifically adapted to transformer models after the more popular, well-known interpretability library, i.e. `captum` [97]. `transformers-interpret` loads the Q&A model from Huggingface, predicting the answer to a question based on a context. When making the prediction, the IG for each input in the text is also calculated. Thus, in addition to the predicted answer, the library outputs the attribution maps for the beginning and end of the answer. Thus, for a Q&A model, two attribution maps are calculated: one for when the answer starts and one for when the answer ends. Note that the stimuli were cleaned of punctuation before processing attribution maps on them. This was motivated by the fact that the models assign separate

Interpretability library	**Downside**
captum [97]	Does not support fine-tuned models from HuggingFace
interpret [98]	Does not support any gradient-based attribution methods
eli5	Does not support any gradient-based attribution methods
interpret-text	Not well-maintained or well-documented

Table 3.12: Libraries we tried for gradient-based attribution generation and their downsides. We eventually chose the `transformer-interpret` library that is well-documented, works with newer models, and has a special wrapper for Q&A models.

attribution scores for all punctuation signs, which would lower our ability to compare human and model ranks, as humans do not see words and punctuation signs separately in the SPR task.

We will now proceed to discuss how we processed the data, as well as what our statistical and descriptive results reveal.

Chapter 4

Statistical Analysis & Results

In the current Chapter, Section 4.1 describes how the data from both humans and models was cleaned and processed, as well as the challenges met. Importantly, in this section, we describe the formula by which we calculate rank differences. Section 4.2 describes scores achieved in performance on FBT, as well as generally alignment scores between the ranks, while Section 4.3 describes the statistical tests conducted and their results.

4.1 Data-processing

4.1.1 Human Experimental Data

After cleaning the experimental data to contain only necessary data for statistical tests, descriptive statistics were used to explore it. This data-processing step first revealed a skewed distribution w.r.t. how participants were assigned to subgroups. Unlike in the pilot study, once published, the human experiment did not distribute participants uniformly, a problem of which the cause was not identified. Because of this, some subgroups have been seen by 22 participants (e.g. subgroup 3), while others have not been seen at all (i.e. subgroup 1). The distribution of participants across subgroups can be seen in Table 4.1, with two subgroups with more participants, 5 subgroups with less than 10 participants assigned, and 1 subgroup with no participants.

Subgroup	2	3	4	5	6	7	8	9
Participants	2	22	20	9	6	12	5	3

Table 4.1: Number of the total of 79 participants assigned per subgroup of stimuli, after the exclusion of three participants due to higher time taken to complete the task. As observed, the third subgroup had the most participants assigned, followed by subgroup 4.

The problem of skewed data distribution was balanced in the statistical analysis by assigning weights w.r.t how much each particular stimulus was seen; see 4.3.1 for more details.

The mean of correct responses of all participants given all items, including fillers, was almost 70%, i.e. 0.68. Out of all 79 participants, only 44 had scores above the mean. We did not exclude participants if they had a lower average of correct responses because we were interested to see how their scores changed with different exposure to questions. Another possible criterion for exclusion

is excluding the participants with a lower mean of correct responses given to Control questions. For Control questions, the mean across stimuli and fillers is 0.79, with only 42 participants scoring above the average. Discarding these participants would result in half the data being discarded. Even more, because some stimuli and fillers were excluded, not all subgroups saw the same number of Control questions; see the next paragraph for more details. For group 4, for example, only the data from two Control questions was kept. This would mean that if the participants from group 4 got one question wrong, they would be excluded, unlike other participants from other groups, who would have more chances not to be excluded, given they were presented with more Control questions. The number of participants that have seen only two Control questions is even bigger, given the uneven distribution of participants across subgroups. However, a criterion for exclusion was the time taken to complete the experiment. On average, participants took 8.04 minutes to complete the experiment, less than we estimated. Out of all 82 participants, 22 participants had longer times than the average. We excluded three participants who had taken longer than 20 minutes.

As aforementioned, three items (two fillers and one stimulus) were excluded from the data because the context in the stimulus was mistakenly part of the first version of the stimuli (i.e., had 55 words), while the responses for the two fillers seemed equally plausible in the context. Regarding the average time spent on each stimulus, Image 4.1 shows the fillers and stimuli that were not excluded had similar average reading times, indicating all the contexts and contextsF were similar in difficulty.

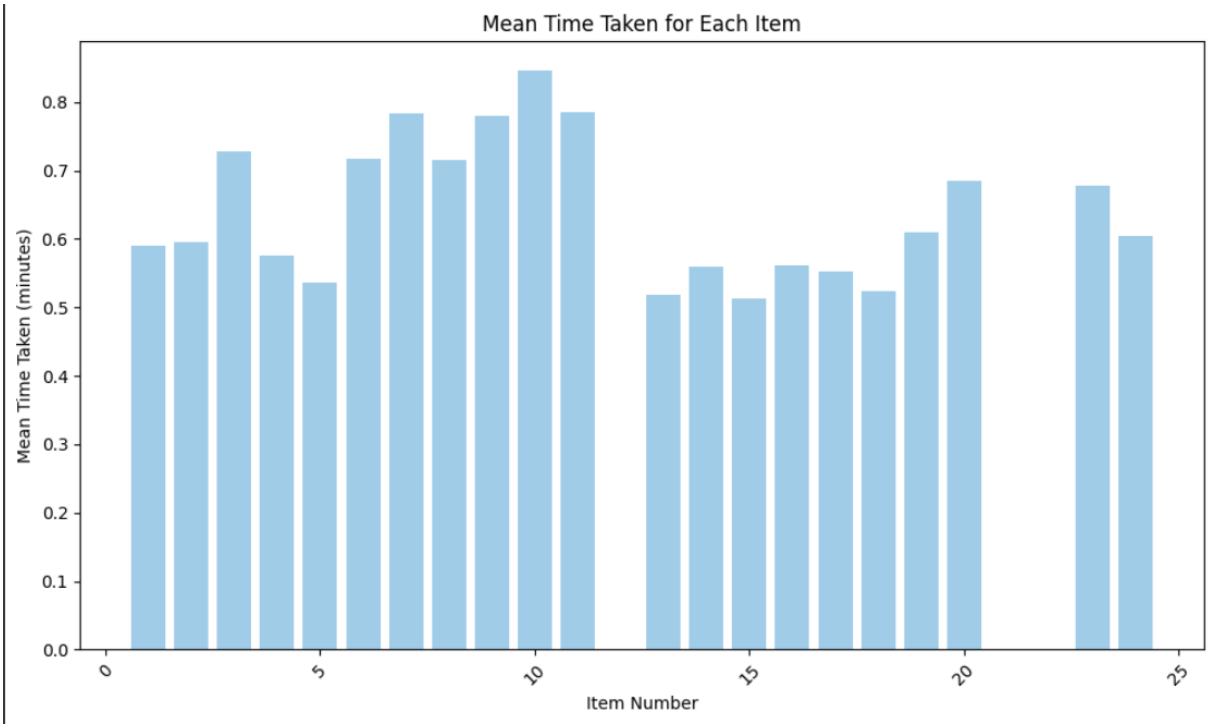


Figure 4.1: Average mean of time participants took per item (filler and stimulus), in minutes. The mean times indicate that contexts in the stimuli and contextsF from the fillers have been similar in difficulty overall, as the values range from 0.5 (half a minute) to almost a minute, but no more than that.

With raw data, we printed the top 10 words with the highest reading times across participants from contexts coming from both stimuli or fillers, appearing before the Control question and after the Agent or Reader question. We used this processing step to manually explore if we could

observe a difference between the words and their reading times, across types of questions, only by looking at raw data.

By manually inspecting the words, we observed that more lexical words appeared in the rank of the top 10 words when proceeded by the Agent or Reader questions, compared to the Control question. Because of this observation, we further classified each word into a lexical or functional word and added that classification to our data structure. The classification was done first by creating a vocabulary of all the words existing in the contexts of the stimuli. These words were further split into two categories: the ones that are always lexical or functional and ones such as the verb 'to be' that can have both lexical or functional uses. The words part of the latter category had their values manually added to the data depending on their use in each context. The dataset contains 75 exclusively functional unique words and 331 lexical words. Given that normally more lexical words compose a sentence, weights will not be used in the statistical analysis to give similar importance to the data as, generally, there is a skewed difference between the two categories. Now we will explain how we processed the attribution maps from LLMs, and how the data was used.

4.1.2 LLMs

Raw attribution data contains the predicted answer of a model for a question and stimulus and two lists of attribution scores (one for the beginning and one for the end of the predicted answer of the model) for each word in the question and context.

The data from attribution maps was firstly cleaned of characters used by models to mark the beginning of text (*[CLS]* in `BERT`, *</s>* in `RoBERTa`), end of sentences (*[SEP]* in `BERT`, *</s>* in `RoBERTa`), subparts of words (*##*) or the beginning of predicted answer ($\dot{G}$ in `RoBERTa`). Note that all punctuation signs in the stimuli were excluded before obtaining the attribution maps, as models would assign separate attribution scores for them; see 3.2.2.

We selected the list of attributions corresponding to the beginning of an answer to mirror the fact that models trained on SQUAD2 receive data only when the answer starts, not where it ends. From the attributions, we extracted each word and its corresponding attribution score. However, `RoBERTa` and `BERT` tokenize words that are not part of their original training vocabulary into smaller word parts. For example, one word that is not part of the original training data for both `BERT` and `RoBERTa` is *non-transparent*. In this case, instead of processing the world as a whole, both models would tokenize it into three different subparts: 'non', 'trans', and 'parent'. As a result, the attribution score for the whole word *non-transparent* would be estimated in subparts, resulting in three attribution scores for the same word. Of all 409 unique words, 190 were out of vocabulary and had attributions for their subparts. To tackle this problem, we took each word that was not part of the vocabulary, found its subparts, summed its attributions, and divided the score by the number of subparts of the word. Subparts of words were identified by comparing the text in the attribution map to the word with the same index in the original stimuli without punctuation. If the attribution text does not match that of the original non-tokenized stimuli, it would be classified as the first subword. Then, we would add the text of upcoming attributions to the original subword until their combination matches the original non-tokenized word. To make the comparison uniform and possible for both models, all words from the original stimuli were lowercased as the model used for `BERT` was uncased. The processed attribution scores were used for three statistical tests testing trends in attribution map scores, the correctness of predicted answers, and human-to-model alignment; see Section 4.3 of the current Chapter for a

complete description of the statistical tests. We will now proceed to describe in more depth how the human comparison was achieved.

4.1.3 Human-Model Comparison

For each word in the context, we calculated the mean reading time across all participants. This option was chosen in favor of other potential strategies of aggregating data across participants because many of the stimuli have been seen by a few participants. In cases with little observations, averaging would work. If we have considered ranking words from the start and then taking the most common rank of one word to be its rank or calculating a median rank for each word, then problems would arise if the stimulus was seen by only two persons.

The mean reading values for each word, along with attribution scores, were ranked into two ranks: one for decisions for humans and one for models. Note that attribution scores were not aggregated across models due to the lower number of models tested compared to the number of participants tested. We took the top 10 words of each rank to find out the difference between the human and model decisions by using the following formula in 4.2. This formula operationalizes the difference between human and model decisions as a difference in the position a word occupies within the rank of highest scores and reading times. If a word has an index of 1 in both top 10s, it will have a difference of 0, meaning it scores similarly in both ranks. The more a word occupies a different rank in the model's top 10, the bigger the difference is. For example, if a word comes first in the human rank, at index 1, and last in the model's one, at index 10, the (absolute) difference will be 9. If a word occupies index 10 in the human rank and 1 in the model rank, the difference will also be 9. If one world is present in the human rank but not the model rank, it scores a difference of 11. The classes 'functional/lexical' were also added for each word present in the rank of highest reading times to statistically test if the classes of words affect the difference of the ranks.

$$\text{Difference} = |IndexHuman - IndexModel| \tag{4.1}$$

Figure 4.2: Formula used to calculate the difference between the top 10 words with the highest reading times and those with the highest attribution values. Absolute values are used to determine the difference. The 'IndexHuman' is the index number of the word in the highest reading rank, and the 'IndexModel' is the index number of the word in the highest attribution score rank. Words that are ranked similarly in their position have a lower difference. If one of the words from the human rank is not part of the model rank, a penalty difference of 11 is added.

4.1.4 Correct responses

For the correctness of responses, we automatically determined the correctness of predicted answers by seeing if they matched the gold standards we specified beforehand. However, as we manually inspected the data, we observed that some predictions were less short but still correct. For example, while the gold standard was 'chargers' the model's answer was 'box of chargers.' Because the model did not choose from multiple choices, as in the human experiment, thereby not being constrained, we thought this type of answer was similar enough to mark it as correct, as it was just more detailed. Thus, we manually inspected the data for such answers, marking

them as correct. Another example is the predicted answer - 'text documents labeled audio documents' and the gold standard 'text documents'. However, we did not consider lengthy predicted answers that contained the gold standard and basically almost all words of the stimuli as being correct. Such an example would be 'text documents labeled audio documents a new intern finds the drive seeing it for the first time he reads the label'. The motivation behind this choice is that unlike in the first case, answers that contain almost all or all words from the context are not just more detailed but are the result of a wrong selection/extraction of the predicted answer.

4.2 Results of descriptive analysis

4.2.1 Humans

As also remarked previously in this chapter, the mean accuracy achieved by humans on the task is 0.68 out of 1. These results suggest that the stimuli tested are more complicated than traditional FBT stimuli and that humans also do not perform optimally on the task. Let us look at the mean of correct responses per type of question. If we consider only the stimuli and not the fillers, we observe from Table 4.2 that the Control question has better accuracy than the Reader or Agent one. The mean of correct responses is also reflected in the plotted version of the table in Figure 4.3. If we look at the mean of correct responses across stimuli considering the belief type, we see that True beliefs have a slightly lower mean than False ones, as observed in Table 4.3. If we look at each type of question and how the mean of correct responses changes if the type of question has a true and a false belief, we observe that participants have higher correct responses on the False belief in the Agent question, as in Table 4.4, whereas in the Reader and Control ones, they have more correct responses in True beliefs. All the means calculated considering beliefs consider only the stimuli, and not the fillers, in the FBT task.

Type of question	**Mean of correct responses**	**Number of observations**
Agent question	0.503	143
Reader question	0.562	146
Control question	0.769	134

Table 4.2: Table representing the mean accuracy across participants for the Agent, Reader, or Control questions. *Type of question* points out which question is measured in terms of accuracy, *Mean of correct responses* illustrates the mean of correct responses obtained for the particular question across participants, and *Number of observations* highlights the number of questions seen across participants and stimuli. For example, across all stimuli and participants, the Agent question was seen 143 times. The numbers come from the fact that in each subgroup, 2 questions from each category were assigned. After the exclusion of some stimuli, the distribution changed in some subgroups. Note that the Control question has not been seen as often, as some participants saw only one Control question in the experiment.

Type of belief	**Correctness**
False belief	0.612
True belief	0.604

Table 4.3: Table representing the mean accuracy across participants for the True and False beliefs. *Type of belief* points out which belief is measured in terms of correctness, *Correctness* illustrates the mean correctness obtained for the particular belief across participants.

We will now proceed with describing the results obtained from the evaluation of models.

Type of Question	Belief Type	Mean Correctness
Agent question	False belief	0.642
Agent question	True belief	0.382
Reader question	False belief	0.513
Reader question	True belief	0.614
Control question	False belief	0.707
Control question	True belief	0.816

Table 4.4: Table of correctness considering each type of question and its belief. The types of questions are reviewed under Column *Type of question*, while the types of Beliefs are reviewed under Column *Belief Type*. The last column represents the mean of correct responses.

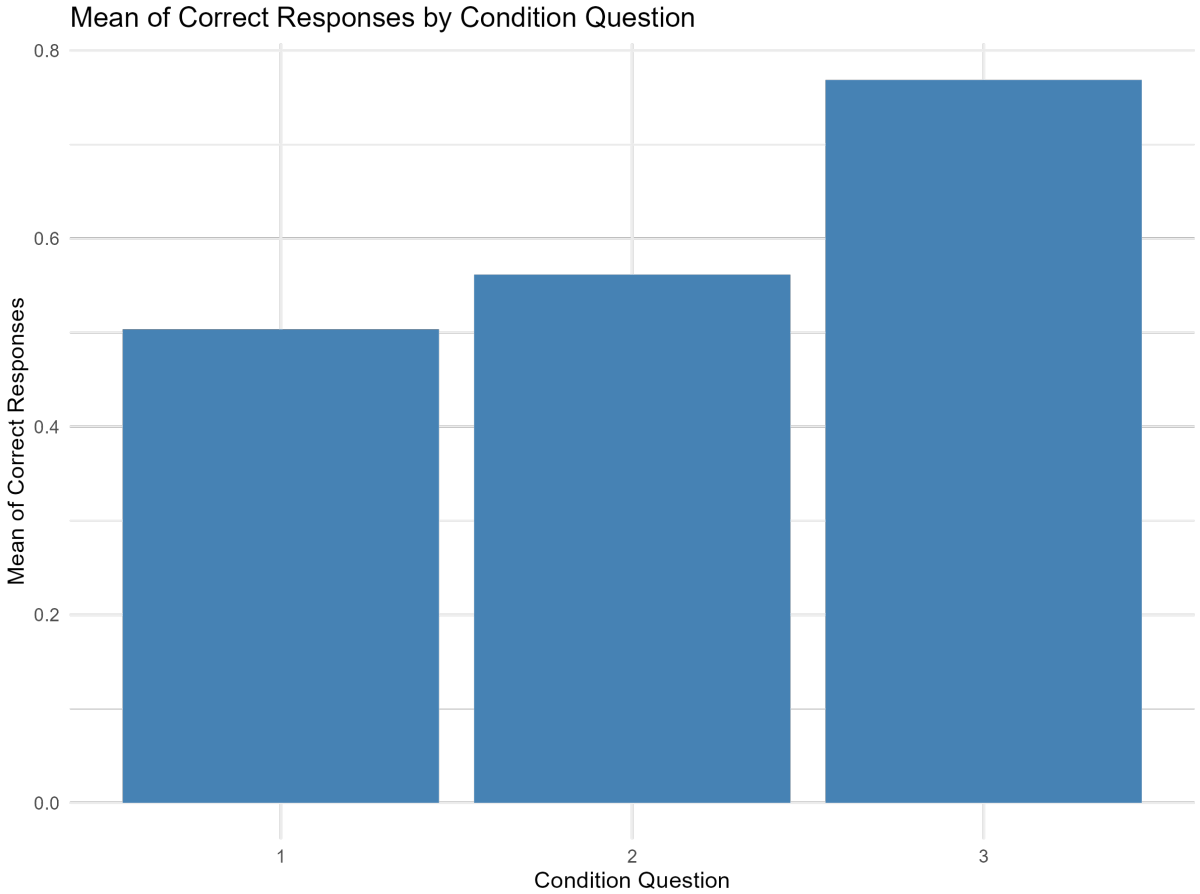


Figure 4.3: Plot of the mean of correct responses per type of question. The type of questions are represented in the plot by the following number: 1 - *the Agent question*; 2 - *the Reader question*; 3 - *the Control question*.

4.2.2 Models

We report the obtained evaluation scores for RoBERTa, shown in Table 4.5 and Table 4.6 for BERT. The names of the models in both tables are determined systematically based on the data on which they were trained, explained in the captions of each table. As can be observed, three general trends are identified. The first one is that training baseline models on train data that they have seen before decreases their performance. For example, both baseline models dropped

their performance by up to 18% when fine-tuned on training data. The second trend is fine-tuning on dev data, which models have not seen before. We see in every fine-tuned model on dev an improved performance on the eval-2k dataset by almost 18%. Lastly, the trend identified in accuracy scores is that fine-tuning models mostly result in improved performance on the FBT task. However, even the best accuracy scores obtained by models on FBT are almost 25% under human performance. The causes of these trends, as well as how these scores relate to previous studies, are debated in the next chapter, namely Chapter 5.

roberta-base-squad2		**Evaluated on**			
	Fine-tuned on	**eval_2k**		**FBT stimuli**	
		EM	**F1**	**EM**	**F1**
Baseline_model	–	76.25	77.89	20.63	23.03
Model 0T	train-8k	59.0	59.71	14.28	15.79
Model 1T	train-I	57.85	58.93	30.95	34.69
Model 2T	train-One	65.3	66.49	25.39	26.80
Model 0D	dev-8k	93.45	94.84	31.74	36.33
Model 1D	dev-I	91.55	93.27	25.39	29.42
Model 2D	dev-One	91.3	93.17	35.71	39.53

Table 4.5: Results for RoBERTa evaluated on eval_2k and FBT stimuli. **Names of models**: *the number* stands for the structure it has been fine-tuned on (0 - no structure, 1 - I+think, 2 - One+think); *the letter* is the origin of the training data (T - subsample from train dataset, D - subsample from dev dataset); *EM*: Exact match scores are correct if the answer exactly matches the gold standard.

bert-large-uncased-whole-word-masking-squad2		**Evaluated on**			
	Fine-tuned on	**eval_2k**		**FBT stimuli**	
		EM	**F1**	**EM**	**F1**
Baseline_model	–	78.5	80.93	29.36	31.66
Model 0T	train-8k	74.25	76.86	38.09	43.37
Model 1T	train-I	63.74	64.81	21.03	25.65
Model 2T	train-One	66.45	67.59	29.36	31.02
Model 0D	dev-8k	97.0	97.98	29.69	33.00
Model 1D	dev-I	96.15	97.31	24.29	27.13
Model 2D	dev-One	96.95	97.94	39.64	44.46

Table 4.6: Results for BERT evaluated on eval_2k and FBT stimuli. **Names of models**: *the number* stands for the structure it has been fine-tuned on (0 - no structure, 1 - I+think, 2 - One+think); *the letter* is the origin of the training data (T - subsample from train dataset, D - subsample from dev dataset); *EM* Exact match scores are correct if the answer exactly matches the gold standard.

4.2.3 Human-Model comparison

If we plot the mean rank difference obtained for all models, we observe that most models score around 8.5, as shown in Table 4.7. In the table, the means of the models show that most fine-tuned models have higher mean rank differences than baseline models, already pointing out that fine-tuning does not align model decisions to human ones. The table also shows that more fine-tuned models do not have the same words in their ranks, as they have bigger counts of a difference of 11. However, fine-tuned models also have more words that occupy the same place in the rank of their decisions as humans, as suggested by generally higher numbers of rank differences of 0. Figure 4.4 is a visualization of the aforementioned table, while Figure 4.5 visualizes the mean of rank difference and the variation. As shown in Figure 4.5, the models have more points clustered around the rank difference of 11.

Model ID	Mean Rank Difference	Count of Rank Diff = 11	Count of Rank Diff = 0	Total Counts
bert_baseline	8.45	382	19	570
bert_0T	8.58	393	20	570
bert_1T	8.17	365	24	570
bert_2T	8.18	366	28	570
bert_0D	8.58	393	24	570
bert_1D	8.28	378	26	570
bert_2D	8.52	394	17	570
roberta_baseline	8.54	387	15	570
roberta_0T	8.50	394	23	570
roberta_1T	8.69	405	21	570
roberta_2T	8.85	413	22	570
roberta_0D	8.50	394	23	570
roberta_1D	8.58	392	17	570
roberta_2D	8.74	399	13	570

Table 4.7: Table showing the mean rank differences across models. Column *Model ID* represents the model's name, while Column *Mean Rank Difference* represents the mean in rank difference per model. This average is obtained by calculating the mean rank difference of all individual words per model. Column *Count of Rank Diff = 11* states how many words had a difference of 11, while Column *Count of Rank Diff = 0* reviews how many words had a rank difference of 0. The total count of words is reviewed under Column *Total Counts*. Note that the names of models are systematically determined after the notation we have previously mentioned when discussing accuracy scores.

4.3 Statistical Tests

We conducted two types of tests: a linear mixed model (lmer), and a binomial generalized linear mixed model (glmer). Note that to avoid confusion, we will refer to them as *tests* or *statistical models* and not just models. When we use only *models*, we refer to language models. The biggest difference between the glmer and lmer, which also motivates our choice, is that one measures the effects of a factor on a continuous variable (lmer), and one does so for a categorical one (glmer). Note that the tests also measure different things: whereas lmers measure how much the dependent variable changes when an effect is added, the glmers measure how much the odds of obtaining a specific category increase by adding an effect. Additionally, one general aspect worth mentioning before describing the statistical results is that only the stimuli were used for all statistical tests, not the fillers. This choice was motivated by the fact that our fillers did not focus on false beliefs but on wishes or needs. Now, we will briefly review some aspects that would help the reader understand the statistical tests.

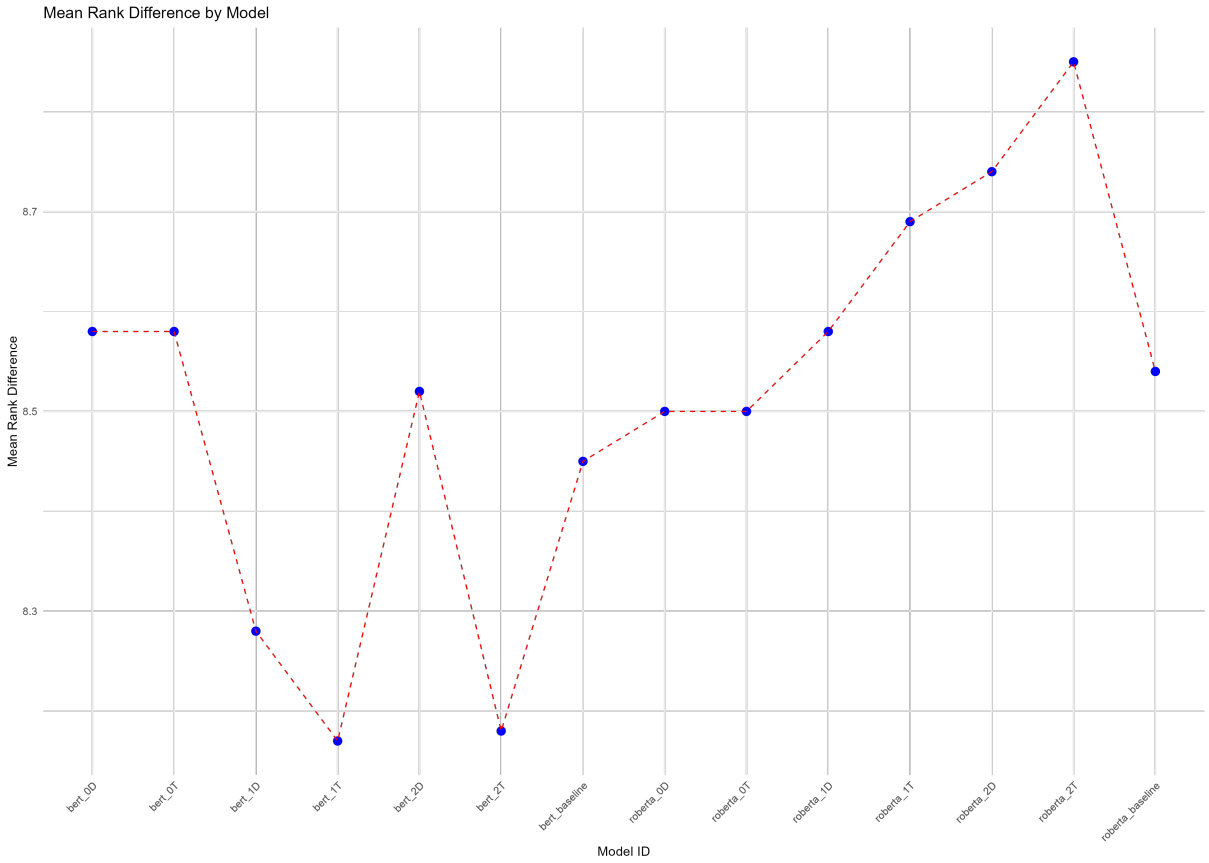


Figure 4.4: Illustration of the means in rank differences for all tested models. The x-axis represents the model name or ID, names of models that coincide with the ones in the accuracy tables in 4.2.2, while the y-axis represents the Mean Rank Difference obtained by the model.

Odds ratios in glmer The results in a glmer, particularly the numbers obtained for how much an effect influences a dependent variable, are in log-likelihoods. Because of this, they can undergo exponentiation to become more interpretable. Thus, we exponentiate each coefficient found significant in a glmer test, and we obtain its odds ratio. If the odds ratio is lower than 1, the category we coded as the 0 category has more chances to appear than one represented by 1 [2]. For example, 0 could be *failure*, while 1 could stand for *success*. Thus, if success is represented by 1, and failure by 0, with an odds ratio of 0.25, failure is four times likelier than success. Similarly, if the odds ratio is above 1, success becomes likelier than failure. An odds ratio of 1 indicates no effect of the factor.

ANOVA tests For some statistical tests, we performed ANOVA tests to tell which effects make a statistical model fit the data better if the potential added effect is not directly crucial to the research questions. The ANOVA reports if one test fits the data better and if that difference is significant. By better fit of the data, we mean that the added effects to the test help explain the data better, leaving fewer unexplained data points. In these tests, we report the so-called AIC and BIC values, which indicate how well the statistical models fit data, with smaller values indicating a better fit.

Random effects Note that for all statistical tests, we tried to set random effects for both stimuli and language models or participants. If a random effect for one of the aforementioned categories is not specified for a statistical test, it means that we tried to add the random effect,

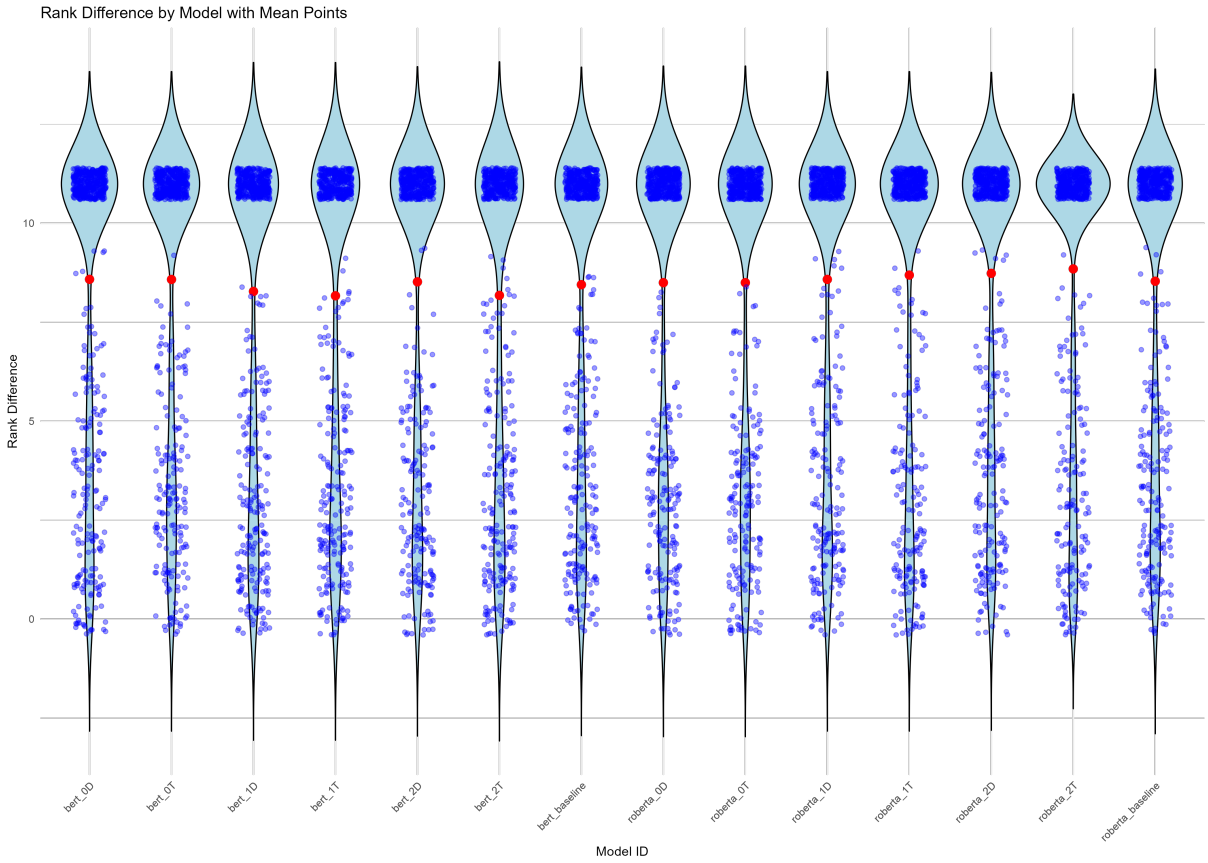


Figure 4.5: Figure illustrating words and their rank difference per model. Words are represented by clusters, which allows us to visualize the rank difference obtained for most words per model. As can be observed, many points are clustered around 11, indicating that most of the top 10 words from human rankings are not part of the models' rankings.

but convergence issues appeared, resulting in its exclusion. Convergence issues appear if not enough data is provided to estimate random effects.

Weights in statistical models Another general aspect we would like to discuss is setting weights. Weights can be used in statistical analysis when certain data points have fewer observations. The weights assigned to data points scarce in observations are bigger than those of data points or stimuli with many observations, thus making the statistical model pay equal attention to both categories. This helps us ensure that the detected significant effects are not because more observations were gained for certain groups. All weights have been calculated similarly: by weighting them as the inverse of how many observations they have, i.e. the more observations for a factor, the smaller its weight. We will now proceed to describe each of the statistical tests and their results.

4.3.1 Human Experiment

We conducted a lmer and several glmer for the SPR experiment. In the lmer, the dependent variable was reading time, and the factors were the type of question, its interaction with the class of the word (lexical/functional), and the type of belief. The test had binary contrasts assigned for the type of belief and class. The type of questions had set a ternary contrast. With this

ternary contrast, we first compared the mean of the Control question to the mean of the Reader and Agent questions and, second, the mean of the Reader and Agent questions. Random effects were set for item number and participants.

In the glmer, we tested if the type of question and belief influences how correct an answer is, and that we ran ANOVA tests to test if an interaction between the type of question and belief fits the data better. Note that the glmer was performed first on all available data (general glmer) but also on subsamples of the dataset to compare more closely how the correctness of response differs when only comparing the Control question to the Reader or Agent question, and not their means. Thus, we will also report results on three subsamples of data: glmer on Control vs. Reader question, glmer on Control vs. the Agent question, and glmer on Agent vs. the Reader question. Note that results on subsamples of data have to be interpreted with caution in the light of the analysis conducted on the whole dataset: the effects detected in the results from subsamples of data do not have to sum up or completely confirm the detected effects in the bigger statistical model, as the statistical models on subsamples will fit the coefficients for smaller samples of data. The statistical models had binary contrasts set for the variables with two levels and random effects for stimuli and participants. In the bigger glmer model, the type of question had three levels, and we set again a ternary contrast that allowed us to compare the Control question with the means of the Reader and Agent ones and the Agent and the Reader question. Note that both statistical models had weights set w.r.t. to each stimulus.

4.3.2 Results Human Experiment

4.3.2.1 Lmer on Reading times

The Control question (β = 29.55, t-score =2.168, p-value $<$ 0.05) as well as the lexical class (β =115.25, t-score = 9.769, p-value $<$ 0.05) were found to be significant. All other effects, including the interaction between class and type of question, were found to be not significant. These results suggest that the reading time of words increases by almost 30 milliseconds if words are not proceeded by a Reader or Agent question. Participants paused for almost 120 more milliseconds for lexical words. The lack of significant effects for the interaction between types of questions and classes of words suggests that lexical and functional words do not differ significantly in their reading times if a question comes after or before them. The lack of significant effects for main factors such as the type of belief or the Reader question suggests that reading time would not differ between types of beliefs or if the Reader question is presented, compared to the mean of Agent and Control Questions.

4.3.2.2 Glmer on Correct Responses

ANOVA tests

For the general glmer, an interaction makes the statistical model fit data better (AIC = 14830, BIC = 14891) compared to the statistical model with no interaction (AIC = 15131, BIC = 15176). For the glmer on the Control vs. the Agent question, a statistical model with an interaction between the type of question and the type of belief fits the data better, as shown by significantly lower AIC and BIC values, i.e. without interaction: AIC = 8234.4, BIC = 8270.3; with interaction: AIC = 7903.9, BIC = 7947. However, for the glmer on Control vs. the Reader question, a statistical model containing an interaction is not more significant than one without,

i.e. with interaction: AIC = 8289.3 BIC = 8332.5; without interaction: AIC = 8287.3 BIC = 8323.3. A similar lack of significance is also observed when comparing the statistical tests with and without interactions for the Reader and Agent question. Thus, the only models that had an interaction were the general and the Control vs. the Agent glmers.

General statistical model

The significant factors found were: Control question (β = -1.04192, z-score = -17.456, p-value $<$ 0.05), Reader question (β = 0.25935, z-score = 4.183, p-value $<$ 0.05), True belief (β = -0.45091, z-score = -2.978, p-value $<$ 0.05), interaction between the Control question and the True belief (β = -1.23127, z-score = -2.796, p-value $<$ 0.05), interaction between the Reader question and the True belief (β = -2.40088, z-score = -2.801, p-value $<$ 0.05). The intercept is β = -0.77782, z-score = -2.756, p-value $<$ 0.05. When exponentiating these results, we see that the odds ratio of getting a correct response when all factors are at their mean is 0.45. For the Control question, the odds ratio is 0.35, which means that compared to the mean of the Agent and Reader questions, getting an incorrect response is more likely in Control questions. However, note that this statistical model compares the mean of the Reader and Agent questions with the Control one and that the mean of the Reader and Agent questions is bigger because they have more data. The odds ratio of the Reader question compared to the Agent one is 1.29, meaning that the correct responses become likelier with the Reader question. The odds ratio for true belief is 0.63. This would mean that the correct answer becomes less likely. The odds ratio for the True belief and its interaction with the Control question is 0.29. For the interaction of the True belief with the Reader question, we have a value of 0.09, meaning that for the Reader and Control questions, the effect of a True belief is smaller. For a visualization of the coefficients, see 4.6.

Control and Agent Questions

The following factors were found to be significant: the Agent question (β = -1.0361, z-score = -9.89, p-value $<$ 0.05) and the interaction between the Control question and the true belief (β = 3.6207, z-score = 16.884, p-value $<$ 0.05). The intercept was β = -1.9304, t-score = -2.925, p-value $<$ 0.05. When we exponentiate these results, we observe the intercept is 0.14 odds ratio. For the coefficient (previously reported by β) of the Agent question, the odds ratio is 0.35. This points to the fact that the Agent question lowers the chance of getting a correct response compared to the Control question. The interaction between the Agent question and the True belief has an odds ratio of 37.36, meaning that the effect of the false belief is greater on the Agent question. This trend is also observed in the descriptive statistical differences between false and true beliefs for the Agent question.

Control and Reader Questions

The Reader question (β = 1.23491, z-score = 14.676, p-value $<$ 0.05) and the False belief (β = -1.23969, =z-score = -14.750, p-value $<$ 0.05) were found to be significant, with an intercept of β = 1.64921, z-score = -3.681, p-value $<$ 0.05. The odds ratio for the intercept is 0.19. The exponentiated coefficient for the Reader question, when compared to the mean of the Control one, is a 0.29 odds ratio, highlighting that the chances of getting a correct response decrease. The false belief decreases the likelihood of having a correct response, as its odds ratio value is 0.28.

Agent and Reader questions

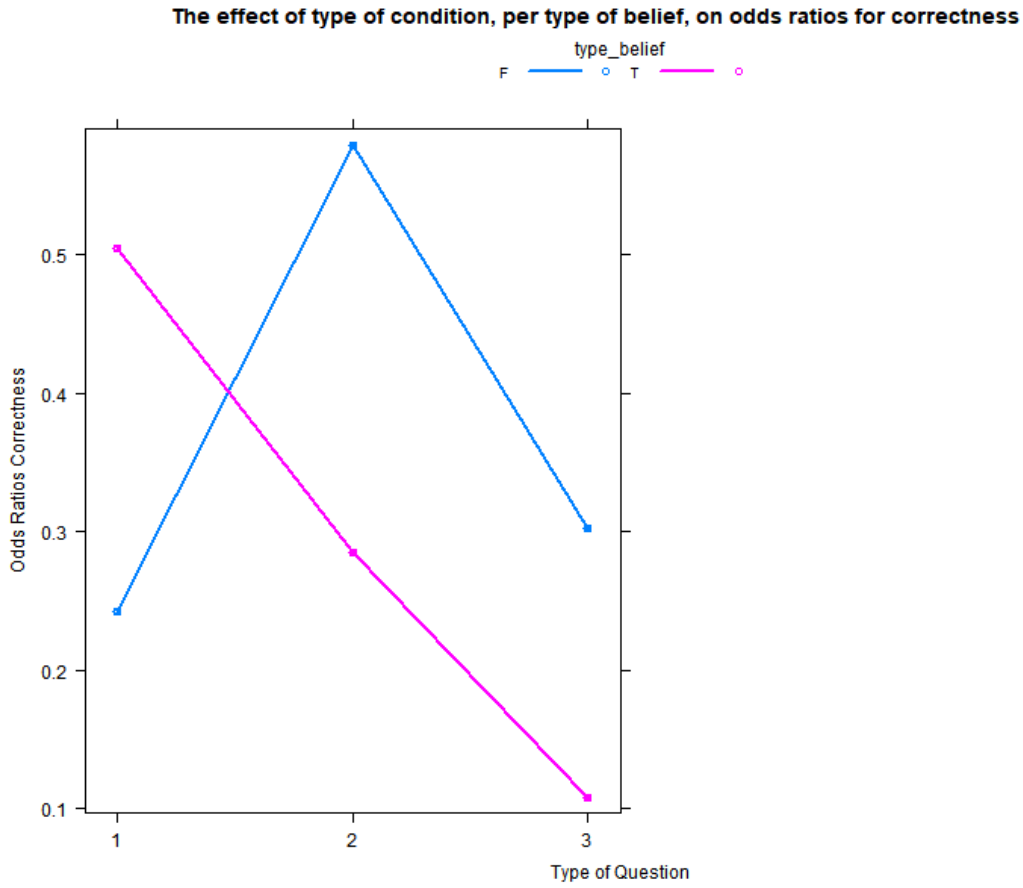


Figure 4.6: Plotted effects in the glmer model for human correct responses. In the plot, the x-axis represents the type of questions and the blue and pink lines represent different types of beliefs. *1* stands for the Agent question, *2* stands for the Reader question, and *3* for the Control question, while *T* stand for the True belief, and *F* for the False belief. The glmer generally shows that when we consider the type of belief, the Reader question under the False belief does the best. The true belief does best in the Agent question.

The Reader question (β = 1.16136, z-score = 5.616, p-value $<$ 0.05) and the true belief (β = 0.39213, z-score = 4.659, p-value $<$ 0.05) were found to be significant. The intercept of the statistical model is β = -0.44653, z-score = -0.712, p-value $<$ 0.05. Thus, if a participant is exposed to a Reader question, giving a correct response is more likely as the value for the odds ratio is 3.19. The true belief also has a positive effect as the odds ratio value is 0.48, indicating an increase in correct responses.

4.3.3 Statistical tests on language models

We conducted two types of statistical tests for language models: glmers testing the probability of getting a correct predicted answer and a lmer testing if scores of attributions, lower or higher, can be linked to any effect.

Glmers

Using glmers, we tested if the correctness of the predicted response is dependent on the following factors: data that comes from a fine-tuned model (also referred to as *factor fine-tuned*), the

type of question, and belief with random effects set for the models. Even though this statistical model was not the one to best fit our data, as shown by several comparisons done by ANOVA between a statistical model with only the fine-tuned factor (AIC = 1719.7, BIC=1735.7), one with no interaction between question and fine-tuning (AIC = 1477.6, BIC = 1504.2), one with their interaction (AIC = 1477.7, BIC = 1514.9) and one with an additional factor of type of belief (AIC = 1479.1, BIC = 1521.6), we still proceeded with it as it was of main interest for our RQs. The statistical model had binary contrasts set for the fine-tuned factor and type of belief and ternary contrasts for the type of questions, with the Control question being compared to the mean of the other 2 ones and the Reader question being compared to the mean of only the Agent one. The test had weights set for fine-tuning to balance the data from fine-tuned language models.

The other glmer tested if the type of structure in the fine-tuning data (I/One/None) affected the likelihood of getting a correct response. The best-fit statistical model found was that with fixed factors for the type of structure and type of question, having lower AIC and BIC values (AIC = 1480.0, BIC = 1511.9) than one with their interaction (AIC =1485.3, BIC =1538.5), one with only a factor for structure (AIC =1722.1, BIC =1743.4), or one with an added effect on belief (AIC =1486.7, BIC =1545.3). However, we kept the interaction between the type of questions and the structure, as well as the factor belief, as they are of interest to our RQs. The test had ternary contrasts set for the type of structure, comparing firstly the non-existence of a specific factor to the mean of the other two structures (I/One), then 'One' to the mean of the other structures (I/None). The ternary contrast for the questions is kept from the other glmer test. Random effects for the statistical models were specified, as well as weights for the types of structures. Because the type of structure was not found to be significant, tests on subsamples of data were not conducted.

Lmers for Attribution Map Scores

A lmer on all-data has the attribution scores as a dependent variable and the type of question, the structure, the class (lexical/functional), the origin of training data (train/dev), and the type of belief as factors. The statistical model had random effects set for items and language models. Unlike previous models where we tested if each added factor improves the fitness of the statistical model, we tested its fitness only w.r.t. having an interaction between the class and the type of question, which did not result in significantly better AIC or BIC values (without interaction: AIC = -28136, BIC = -28040; with interaction: AIC = -28134; BIC = -28022). Thus, we proceeded with the statistical model without an interaction. Ternary or binary contrasts were set depending on the levels of the factor.

4.3.4 Results Statistical Tests on Language Models

4.3.4.1 Glmers Correctness of Predictions

Effect of Fine-tuned on the Correctness of the Predictions

The Control question (β = -2.179947, z-score = -176.410, p-value $<$ 0.05), the Reader question (β = 1.140970, z-score = 67.678, p-value $<$ 0.05), and the False belief (β = -0.143000, z-score = -36.629, p-value $<$ 0.05) were found to be significant. The following interactions were also found to be so: fine-tuned and the Control question (β = 0.665656, z-score = 26.935, p-value $<$ 0.05) and fine-tuned and the Reader question (β = 0.144031, z-score = 4.272, p-value $<$ 0.05).

The intercept of the statistical model was β = 1.120744, z-score = 4.959, p-value < 0.05. Note that the fine-tuned factor was not found to be significant as a main factor. The odds ratio for the Control question, when compared to the mean of the Reader and Agent ones, is 0.11; comparatively, the Reader questions, compared to the Agent ones, have an odds ratio value of 3.12. This points out that having a Reader question increases the chances of getting a correct answer three times, while the Control question decreases those chances when compared to the mean of all other questions. The odds ratio of the False belief is 0.86, decreasing the likelihood of a correct response. Fine-tuning affects the Control question more, with an odds ratio of 1.94, and the Reader question, with a value of odds ratio of 1.15. This would mean that fine-tuning positively affects the likelihood of correct responses more in the Reader and Control questions. Because the comparison between the types of questions is made by using their means, we plot the effects in the glmer for a better look; see Plot 4.7. If we plot the interaction, we see that fine-tuning indeed affects the Control questions more, with them ranking higher in terms of being more likely to get correct responses, see Plot 4.8.

Effect of Structure on the Correctness of Predictions

The following factors were found to be significant: the Control (β = -1.922470, z-score = -289.687, p-value < 0.05), the Reader questions (β = 1.218673, z-score = 131.733, p-value < 0.05), the False belief (β = -0.099180, z-score = -36.631, p-value < 0.05), the interaction between the 'One' structure and the Control question (β = 0.093046, z-score = 6.095 p-value < 0.05), no structure and the Reader question (β = -0.386276, z-score = -22.717, p-value < 0.05), the 'One' structure and the Reader question (β = -0.335107, z-score = -22.717, p-value < 0.05). The intercept is β = 1.296561, z-score = 6.877, p-value < 0.05. The odds ratio value for the Control question is 0.14, indicating it decreases the likelihood of a correct response when compared to the mean of the other two questions. Contrastively, the Reader questions increase the likelihood of correct responses three times when compared to Agent questions, as they have an odds ratio of 3.38. A False belief has an odds ratio of 0.90, pointing out that it slightly decreases the likelihood of a correct response. Having seen no specific linguistic structure correlated with ToM affects less the Reader question having an odds ratio of 0.67. A similar effect is observed when the language model has been fine-tuned on the 'One' structure, and it responds to a Reader question, e.g., odds ratio of 0.71. Having been fine-tuned on 'One' affects Control questions, as indicated by an odds ratio of 1.09. The type of structure or the type of belief was not found to be significant. The effects of the models can be visualized in Plots 4.9, while the interactions can be more closely seen in Figure 4.10.

4.3.4.2 Lmer for attribution scores

The factors that were found to be significant were the lexical category (β = 0.0065123, t-score = 3.731, p-value < 0.05) and the Control question (β = 0.0039162, t-score = 2.088, p-value < 0.05). The type of structure (no structure/One/I), the Reader question, the origin dataset of the data, or the type of belief were not found to be significant. Tests of sub-samples of data for the structures were not performed, given the lack of a significant effect for the type of structure.

4.3.5 Language Model to Human Comparison

We conducted two lmer tests for the human-model comparison. In the first test, we had the difference in rank between human reading times and model attribution scores as a dependent

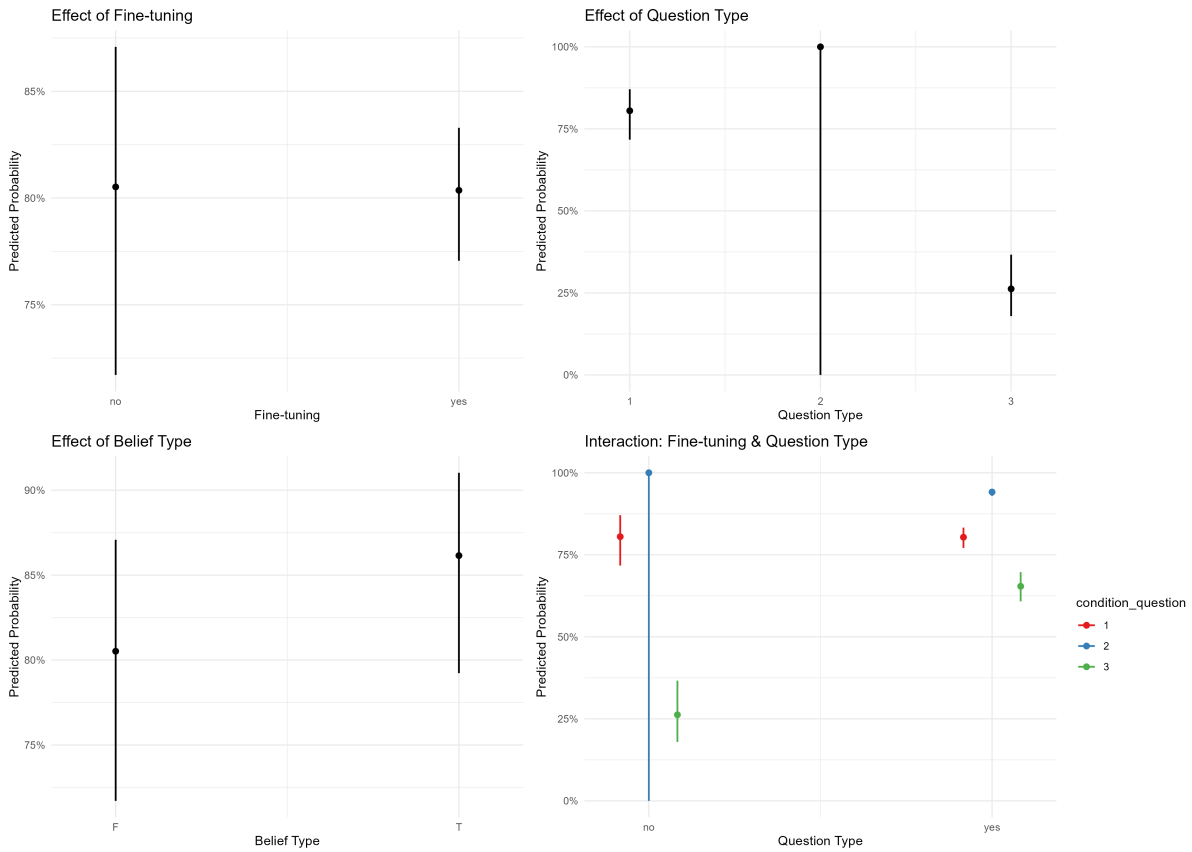


Figure 4.7: Plot of the effects of the glmer on the correctness of predictions in language models. The y-axis *Predicted Probability* are the odd ratios of each factor specified on the x-axis and its effect on the correctness of responses. Here, the odd ratios are scaled into probabilities by R; for example, an odds ratio of 4 means that a correct response is four times likelier than an incorrect one, which would transform to 80%, see [2] for an explanation of how probabilities relate to odds ratios. Note that these probabilities are the same odd ratios mentioned in the statistical test. *Effect of fine-tuning* We observe that the mean probability of getting a correct response from language models that were fine-tuned and ones that were not is very similar. The bigger vertical line for the non-fine-tuned language models represents that the gamer had a larger error in estimating their effect. The closeness in probabilities shows why an effect of fine-tuning was not found to be significant. *Effect of question type* The plot shows that the estimated effects of types of questions differ significantly, with the biggest difference identified between the Control question and the Reader one. In all plots, 1 represents the Agent question, 2 - the Reader question, and 3 - the Control question. *Effect of Belief* The true belief is shown to increase the chances of getting a correct response when compared to the False belief. *Interaction: Fine-tuning & Question Type* the plot shows how fine-tuning affects different types of questions. We again observe that fine-tuning has a bigger effect on Control than Reader or Agent questions.

variable. The fixed factors we tested were if the model was fine-tuned, the type of question, the class of the word (lexical/functional), and the type of belief. Our statistical model had an interaction between fine-tuned and the type of question to test if the type of the question lowers the difference in ranks more for language models that were fine-tuned than those that were not. Contrasts were set depending on the levels of each factor, with the type of question comparing the Control questions with the other two, as well as the Reader and Agent questions.

In the second test, we measured whether the difference in rank is influenced by having been

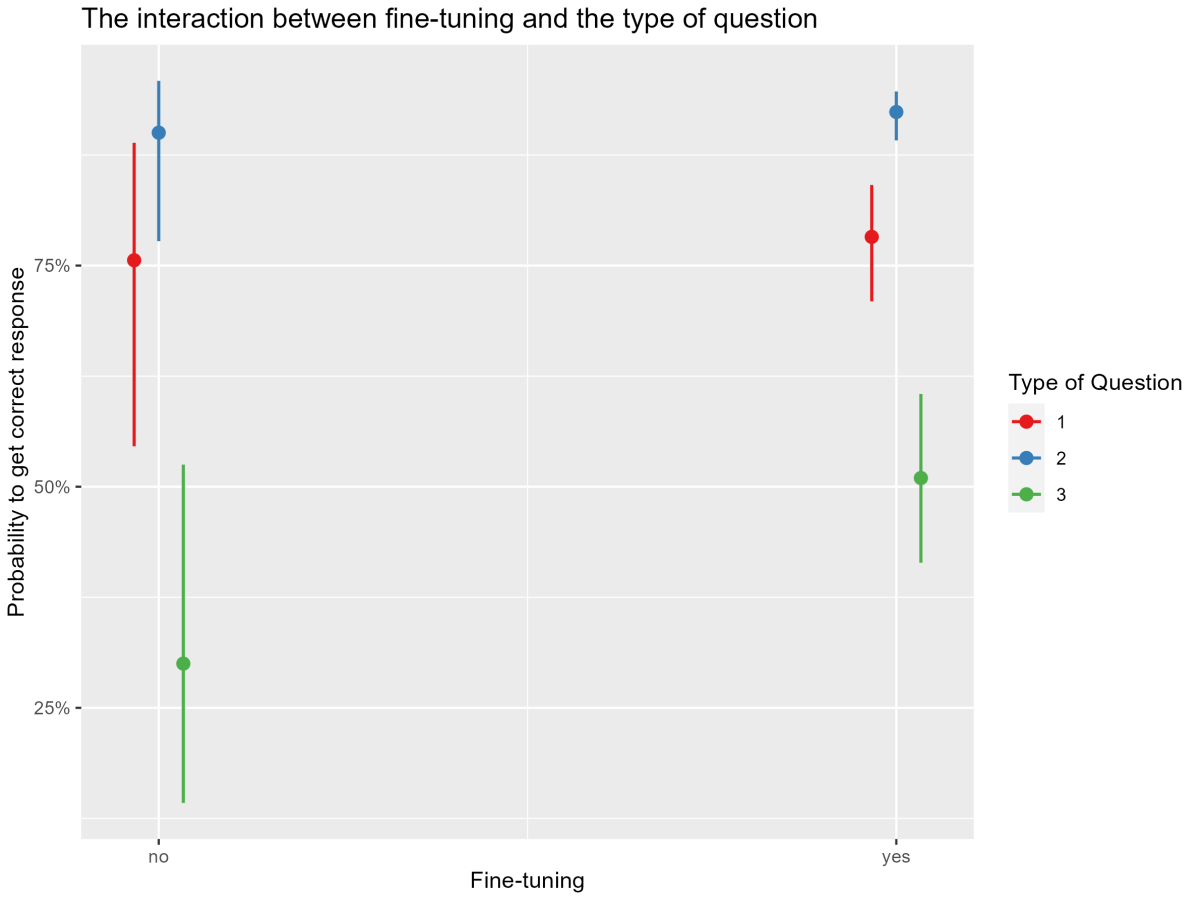


Figure 4.8: Plot of how the fine-tuning factor interacts with the type of question. The x-axis represents whether the language models were fine-tuned or not, and the legend shows the type of question. The y-axis represents the probability of getting a correct response, which is essentially transforming odds ratios in probabilities, as explained by [2]. The plot shows generally that models have more correct responses when exposed to Reader or Agent questions. However, the effect of fine-tuning affects their responses to Control questions more, as the improvement is greater.

trained on a type of structure (One/I/None), the type of question, the class of the word (lexical/functional), and the type of belief. In this statistical model, we had an interaction set between the type of structure and the type of question to observe if language models trained on certain structures are influenced differently by questions containing those structures. Ternary contrasts were set for the type of structure and question. We conducted a separate test from fine-tuning on the type of structure to avoid collinearity issues, as the data points from factors fine-tuned and the type of structures are very similar.

4.3.6 Results Language Model to Human Comparison

Factor Fine-tuned

The only significant factor found was that of class lexical (β = -0.56048, t-score = -4.879, p-value $<$ 0.05). Fine-tuning, the type of question, and the true belief did not have a significant effect on the rank difference. This suggests that despite having more rank differences of 0 in fine-tuned

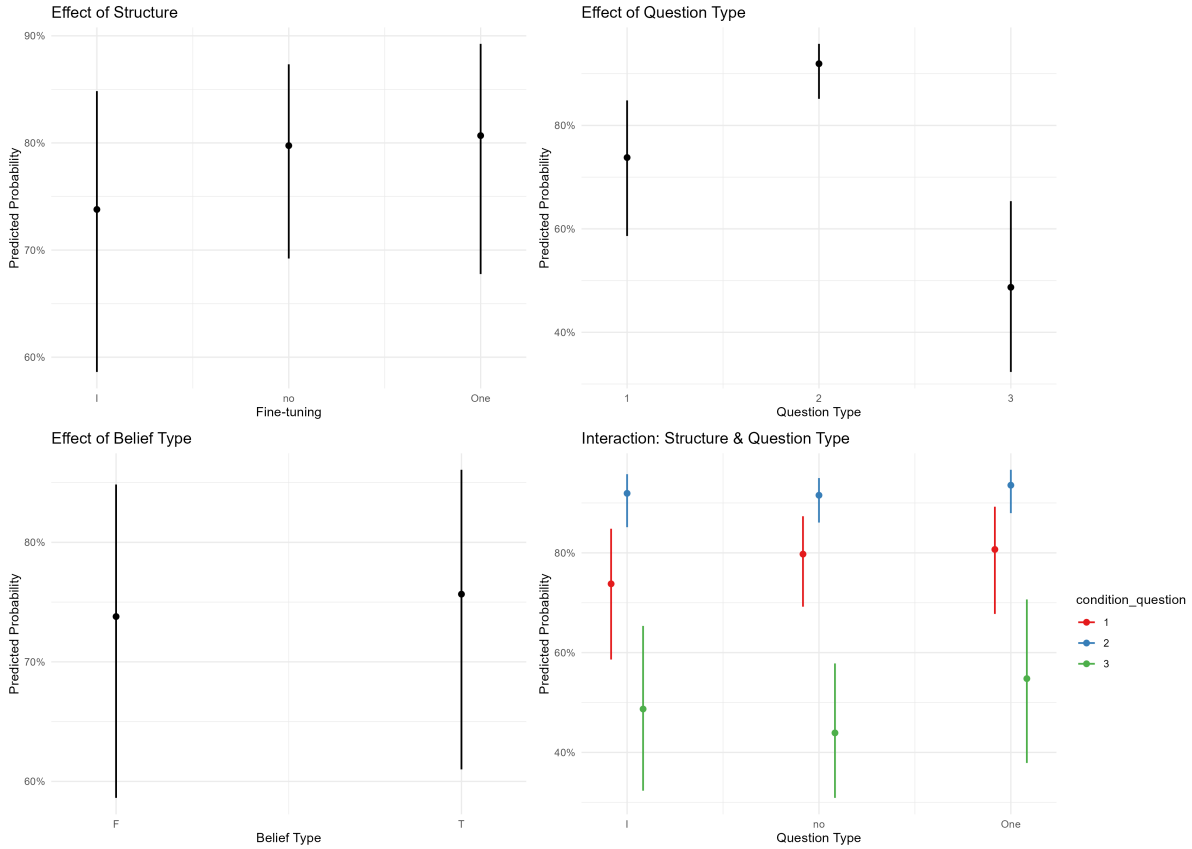


Figure 4.9: Plotted factors from the glmer *Effect of Structure on the Correctness of Predictions*. Note that the y-axis represents the probability of getting a correct response, check [2] for an explanation of the link between odds ratios and probabilities. Plot *Effect of Structure* shows how exposure to linguistic construction can increase the chances of getting a correct response. The means of the types of structures, identified by the dots on each vertical line, do not differ tremendously, which also aligns with the lack of significance found for structures in the glmer model. *Effect of Question Type* plots the effect of Control, Agent, and Reader questions on the probability of obtaining a correct response, with Control questions' positive impact being the lowest. The last two plots show the effects of type of belief in *Effect of Belief Type*, as well as the interaction between types of structures and types of questions in *Interaction: Structure & Question Type*.

models, that effect is not significant enough. The type of question was not found to influence more fine-tuned models than non-fine-tuned ones significantly.

Factor Structure

The only significant factor found was the lexical class (β = -0.56048, t-score = -4.880, p-value $<$ 0.05). Structure One, None, the Control & Reader question, or the True belief were not found to influence the outcome variable significantly. None of the interactions between the type of structure and the type of question were significant, pointing out the fact certain questions do not affect models that were trained on certain constructions differently. See 4.11 for visualization.

Overall Language Models Results

Overall, the lack of significance for fine-tuned models or type of structure indicates that the fine-tuning did not align humans and models. Lexical class reduces the difference between the scores

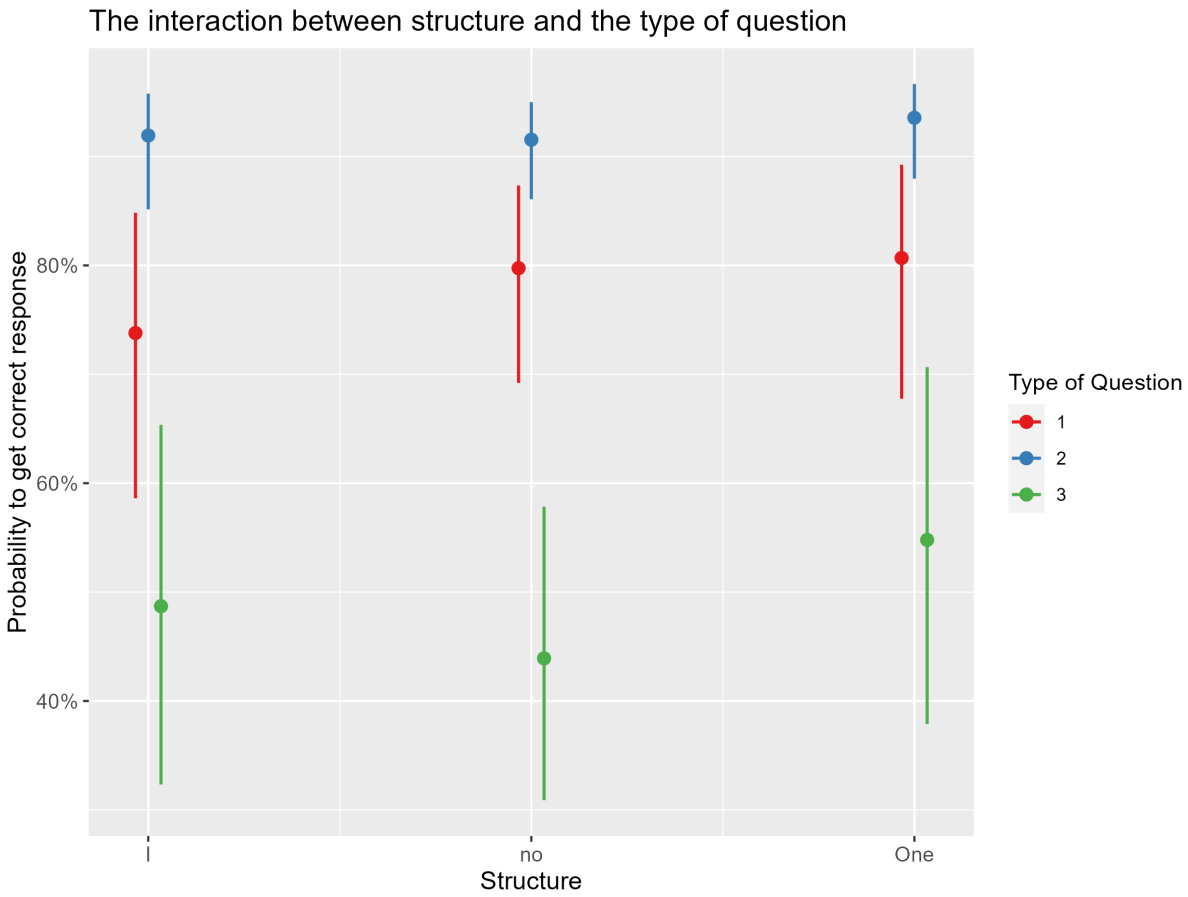


Figure 4.10: Plot of how fine-tuning on specific linguistics structures affects the type of question. The x-axis represents whether the language models were fine-tuned on a particular linguistic structure or not, the legend shows the type of question (1 - the Agent question, 2 - the Reader question, 3 - the Control question), while the y-axis represents the probability of getting a correct response. The plot shows that language models' exposure to certain pronouns or sentential complements matters more if they were exposed to the Reader and Agent questions, with these two having bigger probabilities of getting a correct response. Note that the y-axis represents the probability of getting a correct response, which is essentially transforming odd ratios in probabilities, as explained by [2]

and reading times, which might be because words in lexical class are more likely to be meaningful in answering questions, which might increase their chance of scoring higher in attribution scores and reading times. We will now discuss how these results relate to previous studies and what might have caused them.

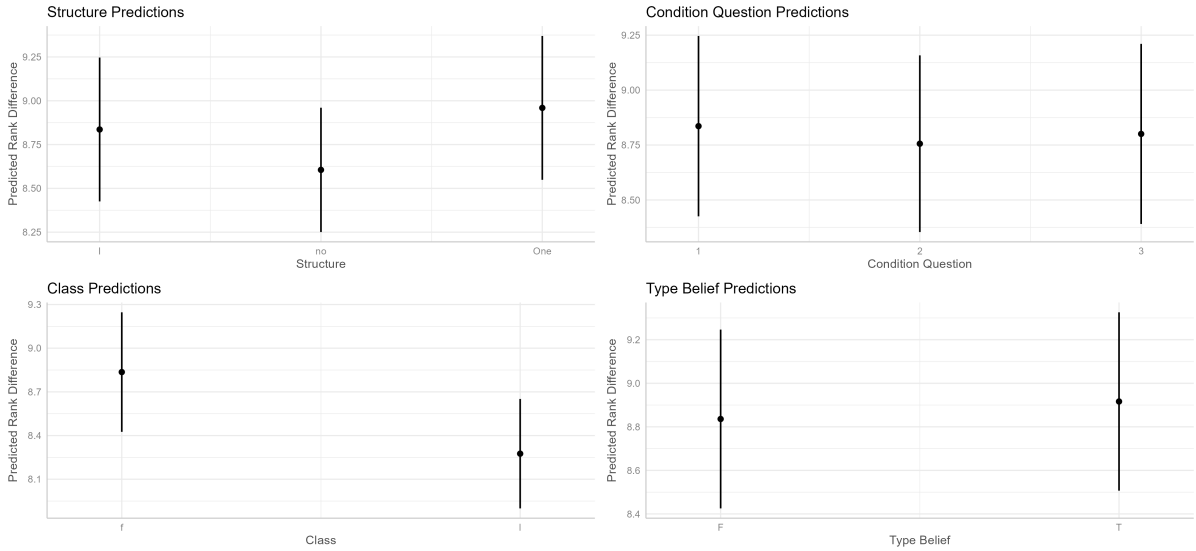


Figure 4.11: Plot representing the effects in the lmer containing the factor *structure*. In each plot, the x-axis is the factor, and the y-axis is the predicted rank difference. *Structure predictions* represents the rank difference explained w.r.t. linguistic structures, showing fine-tuning on no structures results in a slightly lower rank difference. The plot of *Condition Question Predictions* links rank differences to types of questions, whereas the *Type Belief Predictions* does similarly, but for types of beliefs. Note that in the plot for *Condition Question Predictions*, 1 stands for the Agent question, 2 for the Reader question and 3 for the Control one. Both the plots of types of questions and types of beliefs show slight differences in the positions of questions or beliefs. The only plot that shows a difference is that of *Class predictions*, where we observe that the difference in rank lowers when lexical words are considered. The lexical words are represented by the letter *l*, while the functional ones are represented by *f*.

Chapter 5

Discussion

The current Chapter is organized as follows: Section 5.1 discusses the type of question and its effects on correctness in FBT or alignment in humans and models, while Section 5.2 and Section 5.3 do the same for beliefs and classes of words. Section 5.4 discusses general accuracy results for models and reasons for the lack of alignment between humans and our fine-tuned models.

5.1 Types of questions

Humans

Our results show humans achieve around 70% on our task. These results align with previous studies, such as [13], which suggest that humans also perform less optimally on harder stimuli. Our statistical tests and descriptive statistics indicate that the odds of getting a correct response increase when exposed to Control questions. For example, the descriptive mean average is 0.078, while getting a Reader or Agent question decreases your chances of getting a correct response by 0.19 and 0.14 odds ratios. However, note that an improvement in the effect of the Reader question, when compared to the Agent one, is observed, with the Reader question having slightly more correct responses. These observations suggest that there might be a slight influence of priming participants for better FBT performance by means of the 'I' pronoun but that this effect is not a general trend, as participants do not have a similar improving effect when only exposed to sentential complements. Additionally, Agent and Reader questions were read faster, with words in the Control question having higher reading times by almost 40 milliseconds.

The first possible cause of the difference in reading times between the Control question and the Reader and Agent ones would be the position of the questions. It is true that, for example, participants knew what the Agent and Reader questions inquired about before being exposed to the context, which could lead to faster reading times in Agent and Reader questions. When it comes to a possible link between more correct responses and the position of questions, previous studies have indeed shown that asking questions before a task can improve performance, such as in problem-solving [99]. However, the advantage of a question being before or after a text is still debated w.r.t. to reading comprehension, for example, as questions after texts seem as efficient as questions before them [100], also shown in the meta-analysis of [100]. Thus, no difference in the performance of correct responses should have been expected to questions after because they should prove as efficient as the questions presented before. However, the positioning of the

questions cannot be fully excluded as an effect on the difference between questions, as we have no Control questions before the context or Agent and Reader questions after it.

The second possible factor for the difference between Control and Reader and Agent questions is higher processing costs linked to the Reader and Agent ones. For example, [101] show that false beliefs presented earlier in the stimuli have higher processing costs, which result in fewer correct answers as the participants need to remember more things; result also shown in studies such as [48, 102, 103], where working memory interferes with adults' performance or ability to improve on ToM tasks. Because Reader and Agent questions focus on what other agents in the contexts would think, or even more, mention the presence of the reader (i.e. the participant), it would require remembering more aspects of the story. Contrastively, the Control questions only asked what would be in an object central to the story, such as a jar or box.

A small difference was also observed between Agent and Reader questions, with Reader questions having a slightly bigger ratio of correct responses. This effect might have been because the pronoun 'I' would have led to more self-awareness and, consequently, better performance. Thus, our results suggest that there might be a short-term priming effect that improves participants' ToM accuracy. These results are complementary to previous studies that have shown pronouns and sentential complements to be correlated to ToM performance [17–19, 21], thus further supporting our initial hypothesis according to which ToM can be improved, at least w.r.t. humans. However, this effect is small and only observed when pronouns and sentential complements are used together, and the overall difference between the Control and Reader questions still proves bigger. Note that the difference between these two questions has two confounds: the position and the degree of complexity, which future studies should consider to confirm a possible positive effect of pronouns or sentential complements. This shortcoming is discussed in the next chapter. We will now discuss the effects of the type of questions on the models' performance.

Language Models

Models replicate slightly similar trends in the results. Attribution scores are higher for words from stimuli with Control questions by 0.0039162, just like their higher reading times in the human data. However, unlike for humans, such an effect can be immediately pinpointed to how IG is calculated and how lengthier sentences can affect it. It might be the case that in Control questions, the required information is clearer than in the Agent or Reader questions, which can result in the model being more sure about which words weigh more for a good predicted answer. For questions like the Agent and Reader ones, the attribution scores can be more evenly distributed because also the question is less 'clear' w.r.t. what needs to be extracted from the context, which results in words in the context being weighted generally more evenly, as the model is less sure.

When we consider only the effect of fine-tuning, our results suggest correctness increases three times when models receive Reader questions compared to Control questions, where it decreases with an odds ratio of 0.11. Looking at the plot of effects in 4.7, we observe that while there is a bigger error estimation for the Reader question, both the Agent and Reader questions are situated higher in the graph, suggesting they improve the chances of getting a correct response compared to the Control questions. When we consider the glmer model on the effect of being fine-tuned on certain linguistic structures, we observe that the Reader increases the chances of getting a correct response by around three times, while the Control question decreases it. Even more, the plot in 4.10 points to the idea that even if the model has not seen these structures before, the odd ratios for the Reader and Agent questions are higher. These results would suggest that there could be a priming effect for models exposed to questions with certain linguistic constructions.

However, because there are no previous studies specifically testing if pronouns or sentential complements in questions improve the performance of models, our results, which are the first to suggest such an effect might exist, need to be interpreted cautiously. For example, the improved effect might result from the models being less sure of what the Reader and Agent questions ask. Our accuracy results indicate that models still score under chance (under 50%) on FBT tasks. If we make the models more unsure of the question, it would result in them having more evenly distributed decisions (thinking more words are important for the decisions), which would eventually increase the chances of the models giving a response that, because of their overall bad performance, might prove to be correct. Even more, the improvements brought by the Agent and Reader questions are directly correlated to their length, as the lengthiest sentence (the Reader question) brings the biggest improvement, further confirming the possibility of our suggestion that the lengthier the sentence might be, the better the model might perform.

Human-Model comparison

The types of questions were not found to significantly lower the difference between ranks of words in humans and models. Such a result suggests that humans and models do not have more similar ranks if they are exposed to similar questions. However, priming models with certain questions should not change the attribution scores of models obtained from what they learned in the training phase.

5.2 Types of beliefs

Humans

For humans, false beliefs slightly make correct responses more probable, as suggested by our descriptive statistics. Our glmer test also confirms this trend, as the odds ratios for true beliefs point out a decrease in the likelihood of getting a correct response. Depending on the type of question, true belief might result in better accuracy (Reader and Control questions), while in the Agent question, participants exposed to the False belief score better. Our results suggest that false belief can be correlated with better performance. This is unlike what previous studies have shown. For example, [101] test adults on their ability to match pictures depicting true or false beliefs of characters. Their study shows that participants perform much better on stimuli where an unrelated belief about the main story is presented to participants, suggesting false beliefs have higher processing costs.

Models

For models, the false belief slightly decreased the chances of the models getting a correct response. This result is also in line with previous studies, as [24] show the models get more correct responses in True beliefs, i.e. around 70%, than on false ones, around 50%.

5.3 Lexical class

Humans

For humans, lexical words had higher reading times in all tests by 115 seconds. This aligns with previous studies about the length of words and reading times. On average, lexical words are

longer than functional ones, so they would be expected to take longer to read [104] and, thus, to affect reading times more [105]. Because lexical words also contain information related more to the story, it is more than likely that participants paused longer for such words. The reading times of functional or lexical words also do not differ according to the type of questions, meaning they are similar for questions presented before and after the context of the stimuli.

Models

Similarly, the result of the language models suggests that lexical terms weigh more in their decisions by having increased attribution scores by 0.0065123. It might be the case that, as for humans, words with lexical content are more relevant for the model's decisions, given they contain more information. This score can also be linked to the attribution scores being calculated w.r.t. to the question for that particular context in the stimulus, as explained previously. Questions inquire about *what* is the opinion of a character or *what* is in a box, for example, making it more likely for lexical words to be deemed more important as they refer to objects and beliefs. It is also grammatically more expected to pick a lexical word as the answer to a question using *what* than a functional word.

Human-model comparison

When we compare humans to models, the lexical class is the only factor that brings human and model ranks for decisions closer. These results are in line with previous studies such as the one of [42] where humans and models pay attention similarly to lexical classes. This aspect further sustains our previous suggestion that models also regard lexical words as important in making decisions.

5.4 Fine-tuning

Models

The accuracy scores of models show that they perform 20% worse when compared to humans, achieving the best performance when having almost 50% correct responses. Their performance can also be as poor as 25%. These results are in line with the findings of [11, 15] where models perform around 60%, and especially with the findings of [9] that show that on more difficult ToM stimuli, such as the ones we generated, models perform around 30%.

There are also three trends observed w.r.t. fine-tuning, as we remarked in the previous chapter. The first two are a decrease in performance caused by fine-tuning on data from the train sub-sample and an opposite effect of improved performance by fine-tuning on the sub-sample from the dev data. These results further confirm other studies suggesting that exposure to more data leads to improvement [106], which was indicated by our results obtained after fine-tuning on the datasets coming from the dev/validation sub-sample of SQUAD2. For example, [107] suggest larger models use information better than smaller ones, while the study of [108] points to the fact that enlarging the training data has an improving effect even when we consider the same number of learned parameters for a model, with models generalizing better with more data [38]. However, repeating the training data might result in decreased performance, as shown by [109], indicating fine-tuning does not only achieve a positive effect [110]. This is also suggested by the decreased accuracy results obtained from training on data the models have seen before.

The third identified trend is generally improved performance for fine-tuned models on FBT stimuli. We will now use the statistical results to discuss the origins of this improvement further.

Despite the improved accuracy scores, statistical results suggest fine-tuning does not affect the correctness of responses of the models or their attributions scores, a lack of effect which can be seen by fine-tuned and non-fine-tuned models having similar means in the plotted effects of the statistical test in Figure 4.7. Neither does fine-tuning on certain linguistic structures, as observed by the similar means in 4.11. This suggests that the improvement we observed in accuracy scores is not significant enough statistically to be attributed to a model being fine-tuned generally or fine-tuned specifically on pronouns and sentential complements. This is also observed in accuracy scores, as fine-tuning on structures does not always improve accuracy; for example, BERT performs worse in the FBT task when trained on the 'I' pronoun. The fact that some models improved also rules out that our fine-tuning hyperparameters might have destroyed the model's performance, as we have models that improved from solely fine-tuning and fine-tuning with certain structures.

However, interesting interaction effects were found between fine-tuning or structure and the type of question. Fine-tuning of the models interacts with the Control question, i.e. impacts it more positively than the Reader or Agent question, significantly improving the performance of fine-tuned models on Control questions, as observed in 4.8. In the same model, we also see an improvement for the Reader question; the graph shows one for the Agent one as well. However, the bigger effect for Control questions might be because the models exposed to more data become better at traditional extractive question answering, which is basically what our Control questions target. This is observed by looking at F1 answers in accuracy scores. For example, in our results, models trained on train data indeed have worse performance on the eval-2k dataset sampled from SQUAD2. However, the performance they get after fine-tuning on dev data is majorly bigger, scoring close to even 98%. This improved accuracy might lead models to get more accurate responses to the Control question after being trained on dev data, which would improve their overall performance when all models are considered for statistical analysis. This would also explain why fine-tuned models do not record a similar effect if they are just exposed to the Agent and Reader questions, as those test ToM besides extractive Q&A, which remains almost the same as suggested by the statistical results.

These observations are in line with the findings of [111]. The authors test if a model reacts differently to out-of-domain data that was not part of the fine-tuning dataset after fine-tuning. Again, note that testing on out-of-domain data is essentially testing a model on a task it has not been directly fine-tuned for, exactly like we did in our FBT task. The study of [111] argues that a fine-tuned model has similar answers to a non-fine-tuned one on out-of-domain data. These previous remarks are, again, in line with our results that suggest there is a bigger influence of fine-tuning on the Control questions, as the main fine-tuning task was extractive Q&A. This is related to previous studies where fine-tuning for a task improves performance by creating more distance between the labels for classification, as shown by [110]. [111] as well as [110] look at how similar are representations for BERT before and after fine-tuning. [111] compare models fine-tuned exactly on the SQUAD dataset. They look at the representation spaces of questions and conclude that the last layers have the most differences, with [110] concluding that the difference would appear more in earlier layers. Regardless of their results, the key idea of both studies would be that models manage to become better in fine-tuning by creating more space between different classes or labels. This is why out-of-domain fine-tuning would be harder, as it might be hard for models to create more distant spaces in ToM, for example, if the dataset does not target that task specifically. Our results on attribution scores also reinforce these findings: fine-tuning

on questions with 'I', 'One', and 'think' did not result in different attribution maps for FBT stimuli, pointing out again that the models act the same as baseline models on out-of-domain questions and stimuli, which are our FBT stimuli. These results also suggest that while humans can be trained and prompted to do at least slightly better in ToM with these structures, such an effect seems less likely for models. Thus, the current results suggest that fine-tuning on certain linguistic constructions correlated with ToM improvement in humans does not improve models' FBT performance. Even more, our results suggest that the decisions of LLMs and humans are not similarly linguistically informed, as an improving effect would have been likely to appear if the models had learned certain linguistic cues to be linked with agents' beliefs. We will now proceed to discuss how close the ranks of important words are in humans and LLMs.

Human-Model comparison

Firstly, our results point to the fact that, generally, models do not align well with human decisions, as shown by the scores in Table 4.7. These results are contrary to what [41] have shown about LLMs replicating human tendencies or values in simulating brain scan activity or associations of unrelated concepts from different domains. [42] study how close attention maps are between human and models in a text classification task, showing that models also select all the words selected to be important by humans. We show an opposite effect, i.e. that they are not similar, indicated by the differences in the ranks. Our results might be different due to our methodology. Unlike [42], we did not ask humans to rank what they find as important for the task, which can introduce a bias of them selecting what they think they are supposed to find as important. Because we have directly processed their reading times, this might have led to signaling more of the difference between humans and models.

Another study that suggests humans are similar to models is [44], which shows that the first important word in the decisions of humans in a text comprehension and classification task is also part of models' words deemed as important. Our results do not suggest this, if we just manually check the data, we immediately observe that the first word important in the human rank for the first three stimuli is not even present in the model rank. On the other hand, our results suggest that models and humans can focus on similar words, but not always, as suggested by [66]. Thus, unlike in previous studies, the difference between humans and models is bigger, with no improving effect caused by pronouns or sentential complements. The differences in the human and model ranks might be due to the lower performance of models in the FBT task. In all cited papers, models do well on the researched tasks, unlike our FBT task. Note that alignment to humans has proved to indicate better performance in the past in other tasks [43]. Our results partially reinforce those observations as they suggest that the models that perform badly are not similar to humans. However, our results do not suggest causality between worse performance and difference in ranks between humans and models. Thus, the only partial reinforcement of the results shown in [43].

Our results further indicate that fine-tuning generally was not found to lower differences between humans and models w.r.t. the ranks of words, nor was fine-tuning of a specific structure. These overall results suggest that reading times and attribution scores are not made more alike by fine-tuning, which is also sustained by a lack of significant effect of fine-tuning on attribution scores. Unlike in [64], which have obtained alignment by fine-tuning on human annotation data, or [65] that obtained good correspondence between LLMs and humans in planning the schedule of one day, we did not obtain human to model alignment. As previously remarked, our lack of alignment might be caused by training for out-of-domain performance. The failure to align models to humans can also be due to a lack of human-informed data for fine-tuning. The studies of [64, 65] suggest that alignment might be achieved if human data is provided, for example, as

[66] have shown the attention maps of human and models have a higher correlation if the model is exposed to human eye-tracking data. The same improved alignment was found in [67], where models have been trained on attention maps from humans. The fact that pronouns or sentential complements would be out-of-domain for ToM in models suggests that the models have not, at least, learned them as cues relevant to ToM.

However, aligning by using human data might prove to be eventually very costly [112] as, for example, in our case, we would need data for human decisions, which would imply experimental collection at a larger scale. To account for such a problem, alignment might be achieved by ranking the already learned parameters of models by preference functions of human values, as suggested by [112]. Thus, though we have not achieved alignment, our human data can be used for human-preference alignment, for example, as was done in [113].

Chapter 6

Conclusions and Limitations

6.1 Conclusions

The current thesis tested both humans and LLMs regarding their decisions in the FBT task, how alike they are, and if they can be made more similar. To investigate their decisions, we compared two ranks: a rank obtained from the highest reading times from an SPR task and one from attribution scores from language models. To investigate if they can be made more similar in their decisions, we used the pronouns 'I', 'One', and the sentential complement 'to think' to prime participants or fine-tuned models. We have also created new FBT stimuli for the tests, in line with the stories from previous datasets that were found to be difficult [19] for both humans and language models. We created the stimuli by synthetically generating similar stories to the original dataset from [1] with GPT-4, which were checked manually for plausibility and later modified to have similar lengths. Our results show humans perform around 70%, ranking better than LMMs that perform at most around 50%. Human-model alignment is low, around 8.5, indicating that, on average, the difference in the index a word has in ranks from humans to models is almost 9. Thus, our results suggest that humans and language models do not base their decisions on the same words and that fine-tuning models on certain linguistic expressions (pronouns or sentential complements) does not improve their FBT performance or alignment with humans. Contrastively, humans perform just slightly better when primed with questions containing 'I' but not as well as when responding to Control ones. We will now briefly go over the contributions brought by the current thesis, considering humans, models, and their comparison more closely, after which we will discuss the limitations of the current work.

The human experiment

Concerning the study's human experimental design, we presented an SPR design in which questions are presented before the context of the stimuli. This choice makes participants pause more for words important to answering the question, signaling their decision. We primed participants by presenting them with questions that had pronouns or sentential complements in various degrees.

Our results show that humans have overall improved accuracy when exposed to Control questions, almost 80%. This difference might have resulted from the decreased difficulty of the Control questions. In comparison, the mean of correct responses for Agent and Reader questions is around 50%, with Reader questions having more correct responses. The slightly improved effect

of the Reader question might have been because of the exposure to 'I'. Our work is the first to suggest that such a short-term improvement effect might exist for humans on ToM tasks. Lexical words also rank high in human reading times, which further confirms the effect of length of words on reading times in SPR experiments.

The language model experiment

We took the SQUAD2 dataset and modified it to create two new datasets for fine-tuning that rephrased questions to have either 'I+think' or 'One+think' constructions. We have done this with data that the language models have seen before (train data) and with one they have not (validation data), resulting in 4 new datasets. The modification was obtained by prompting GPT-4 to rephrase all questions using two-shot prompting. Afterward, we fine-tuned with them two types of models (`BERT` and `roBERTa`) and obtained 8 fine-tuned models on linguistic constructions that were previously correlated with better ToM abilities, and 4 models only fine-tuned on data with no specific structures.

Our accuracy results show that LLMs perform better on a subsample from the SQUAD2 dataset if fine-tuned on data from the dev dataset, reaching almost 98% F1 scores in some models. An opposite trend is observed for training on train data as performance decreases. On FBT, the models' performance ranges from 16% to almost 45%. The statistical results suggest fine-tuning, or fine-tuning on certain structures, does not improve the probability of models getting a correct response in FBT. However, it might improve its overall extractive Q&A ability, which results in better performance in Control questions. Thus, the slightly better performance fine-tuned models obtained on FBT might be due to its heightened extractive Q&A ability to respond to questions that do not necessarily concern ToM (the Control question). Language models also have greater chances to respond correctly to Reader or Agent questions, which might be because the models assign attribution map scores more evenly across words, as Reader and Agent questions are longer. Models have higher chances to correctly predict answers in True beliefs, a result observed by previous studies as well.

The human-model comparison

We presented a formula to quantify alignment between attribution scores and human reading times considering the highest-ranked words. In our formula, we calculate the absolute difference between how high a word is ranked in humans and model decisions, from 0 to 11, with lower differences indicating a better alignment. Our results suggest models and humans rank words very differently, with around 70% of the words in the human rank not being present, on average, in models, and a mean rank difference of 8.5. This low alignment is unlike in other tasks [39, 42]. We have also shown that fine-tuning with pronouns and sentential complements does not result in out-of-domain alignment, which in our task would be closer ranks to humans in FBT. Note that closer ranks have proved in the past to improve the performance of models on various tasks [43].

We will now proceed to discuss the limitations of the current study.

6.2 Limitations and Future Research

The first limitation pertains to the human experiment: the experimental data is skewed, as some stimuli in some conditions were seen by many persons, while others have not been seen at all. This was partially accounted for by weighting the number of observations for each stimulus

in the statistical tests we conducted for human data. However, when comparing humans to language models, we discarded stimuli due to a lack of data from human participants, which future studies could consider testing. While presenting questions before the text of stimuli is a good idea generally, we have not presented Agent or Reader questions after them. Because of this, we cannot fully attribute the improved FBT scores of humans after the Reader question to linguistic constructions or explain the bigger performance on the Control question due to fully its low complexity. Thus, future studies need to be conducted in which Control and Reader questions are also presented before the context of the stimuli.

W.r.t. language model fine-tuning and testing, our study is limited considering the type of architectures we used, i.e. only encoder-based models, and the models' size. A future improvement could be testing newer and larger models with the current framework. Additionally, the data for fine-tuning might not have been enough to see an improving effect, which could be a path to explore in future studies that would use our methodology to generate synthetic new modifications of current existing datasets with LLMs.

This also stands to the number of stimuli we tested; while not necessarily a limitation, it is certain that more diverse stories could provide a less-biased picture of models' performance, and our current dataset and generation prompting strategy can be used to enlarge the FBT benchmark we currently provide. One thing we did not do was to prompt the model with a higher temperature to produce more varied and diverse scenarios for the FBT stimuli, which can also be explored in future studies. Note also that future studies could consider making the stimuli more similar. While all contexts of the stimuli have the same length, they are not similar w.r.t. to the part of speech that comes at a certain place in a sentence. This means that sometimes the second word, for example, can be an adjective or a noun. This can be damaging for both the SPR task and the attribution scores, as across contexts, the reading times or attribution scores can vary a lot, considering that maybe a sentence has more nouns or adjectives than functional words. This can result in noisy data, affecting the true impact of different effects on the statistical analysis. Thus, future studies could consider improving this divergence within stimuli.

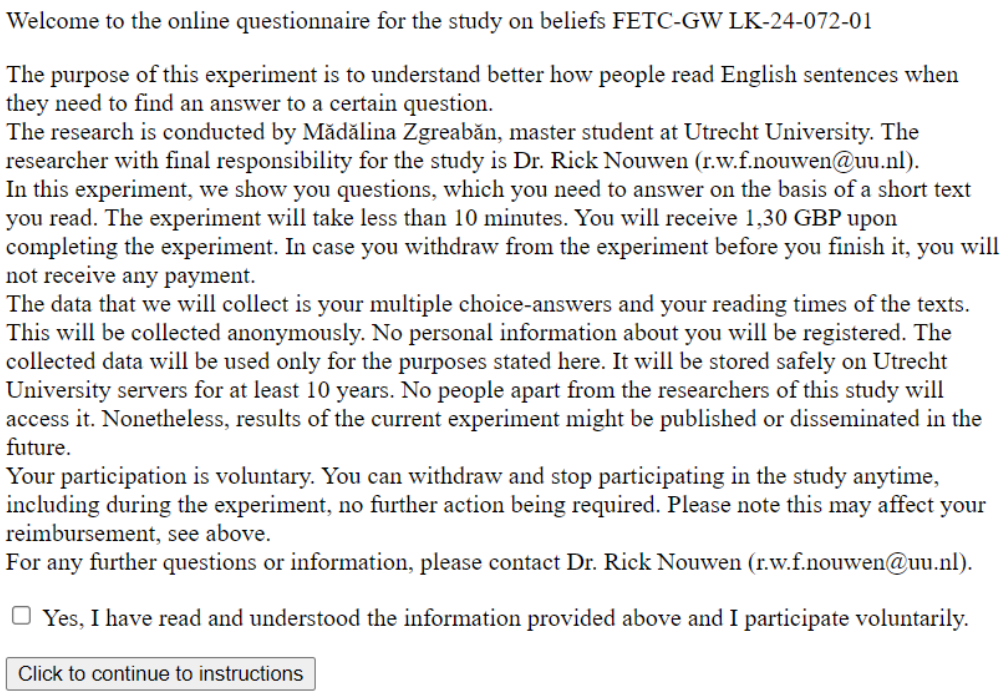
When testing models on the task, several confounds could be accounted for in the future. For example, [11] have shown that models tend to pick the last object mentioned in the prompt. Despite this potential aspect that could explain many of the models' choices, the current study did not mainly test for it as it was out of the domain of our research questions. Future studies can consider integrating this aspect in the testing phase and the analysis of models. Another aspect that could concern future studies is how self-aware the tested models are. For example, [114] test how self-aware LLMs are by their uncertainty shown to unanswerable questions. In their study, unanswerable questions are defined as being scientifically debated, requiring subjective answers, or having more than one possible answer. The uncertainty of the models' answers was calculated by their similarity to sentences with uncertain meanings. They show that bigger models and those trained using instruction tuning are more self-aware. Considering the `davinci` series of models, most score around 50%, while GPT-4 scores around 75%. Even then, GPT-4 scores less than humans by almost 10%. As we have previously discussed, pronouns and their effect on alignment were studied because they are correlated with heightened self-awareness, which is linked to ToM. Thus, it might be the case that some of the results of our models could be linked to how self-aware they are, which might explain the lack of improvement in the FBT task.

Lastly, fine-tuning for model improvement or alignment can be done using the human reading times in the current study, which might result in better performance and similarity to human behavior.

Appendix A

Appendix A

A.1 The human experiment


Welcome to the online questionnaire for the study on beliefs FETC-GW LK-24-072-01

The purpose of this experiment is to understand better how people read English sentences when they need to find an answer to a certain question.
The research is conducted by Mădălina Zgreabăn, master student at Utrecht University. The researcher with final responsibility for the study is Dr. Rick Nouwen (r.w.f.nouwen@uu.nl).
In this experiment, we show you questions, which you need to answer on the basis of a short text you read. The experiment will take less than 10 minutes. You will receive 1,30 GBP upon completing the experiment. In case you withdraw from the experiment before you finish it, you will not receive any payment.
The data that we will collect is your multiple choice-answers and your reading times of the texts. This will be collected anonymously. No personal information about you will be registered. The collected data will be used only for the purposes stated here. It will be stored safely on Utrecht University servers for at least 10 years. No people apart from the researchers of this study will access it. Nonetheless, results of the current experiment might be published or disseminated in the future.
Your participation is voluntary. You can withdraw and stop participating in the study anytime, including during the experiment, no further action being required. Please note this may affect your reimbursement, see above.
For any further questions or information, please contact Dr. Rick Nouwen (r.w.f.nouwen@uu.nl).

☐ Yes, I have read and understood the information provided above and I participate voluntarily.

Click to continue to instructions

Figure A.1: The consent letter on the first page of the experiment.

A.2 Stimuli, variants of response, correct answers

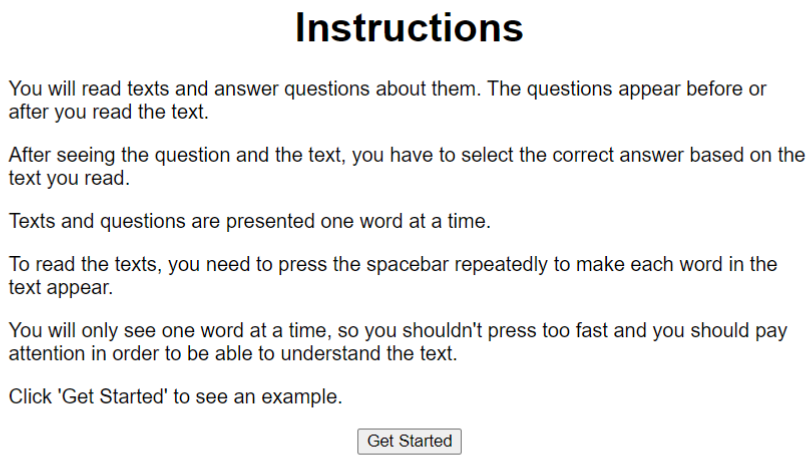


Figure A.2: The instructions on the second page of the experiment.

stimulus	correct_answer	option 1	option 2
In a grocery store, there is a non-transparent jar filled with honey. An employee wrongly labeled it 'wheat' instead of 'honey'. A customer reads its label, opens the jar, looking inside. The customer closes it.	honey	wheat	honey
In a grocery store, there is a non-transparent jar filled with honey. An employee wrongly labeled it as 'wheat' instead of 'honey'. A customer sees the jar, reads its label, and decides to buy it.	wheat	wheat	honey
In a grocery store, there is a non-transparent jar filled with honey. An employee wrongly labeled it 'wheat' instead of 'honey'. A customer reads its label, opens the jar, looking inside. The customer closes it.	honey	wheat	honey
An encrypted file labeled 'recipes' contains research notes. A data recovery specialist is hired to open it. After reading the name of the file, the specialist fails to open it. So, the specialist stops trying.	recipes	recipes	research notes

Stimuli	Correct answer	Option 1	Option 2
An encrypted file labeled 'recipes' contains research notes. A data recovery specialist is hired to open it. The specialist is informed of its contents, and opens the file, inspecting them. The specialist closes the file.	research notes	recipes	research notes
An encrypted file labeled 'recipes' contains research notes. A data recovery specialist is hired to open it. After reading the name of the file, the specialist fails to open it. So, the specialist stops trying.	research notes	recipes	research notes
A non-transparent sealed shipping container containing rice arrives at a port. Documents state it contains rice. An officer checks the documents, looks at the container, and opens it, looking inside. The officer then closes it.	rice	rice	wheat
A non-transparent sealed shipping container containing rice arrives at a port. Documents state it contains wheat. An officer checks its documents, looking at the container, but fails to open it. So, the officer stops trying.	wheat	rice	wheat
A non-transparent sealed shipping container containing rice arrives at a port. Documents state it contains rice. An officer checks the documents, looks at the container, and opens it, looking inside. The officer then closes it.	rice	rice	wheat
A quarantine inspector receives a transparent bag, labeled as 'swabs', with swabs. The shape of the receptacles looks like those for blood samples. The inspector fails to open the bag. So, the inspector stops trying.	blood samples	blood samples	swabs
A quarantine inspector receives a transparent bag, labeled 'swabs', with swabs. The shape of the receptacles looks like those for blood samples. The inspector opens the bag and inspects its contents The inspector closes it.	swabs	blood samples	swabs

Stimuli	Correct answer	Option 1	Option 2
A quarantine inspector receives a transparent bag, labeled as 'swabs', with swabs. The shape of the receptacles looks like those for blood samples. The inspector fails to open the bag. So, the inspector stops trying.	swabs	blood samples	swabs
In the kitchen, there is a transparent red box with cookies, labeled 'crackers'. Alex is looking for crackers. He knows his mom put a box of crackers in the kitchen. So, he takes the box.	cookies	cookies	crackers
In the kitchen, there is a non-transparent red box with cookies, labeled 'crackers'. Alex is looking for crackers. He knows his mom put a box of crackers in the kitchen. So, he takes the box.	crackers	cookies	crackers
In the kitchen, there is a transparent red box with cookies, labeled 'crackers'. Alex is looking for crackers. He knows his mom put a box of crackers in the kitchen. So, he takes the box.	cookies	cookies	crackers
At a crime scene, there is a closed crate containing gold bars, labeled as 'gold bars'. An investigator thinks it contains counterfeit currency. Despite trying, he cannot open the crate. So, the investigator stops trying.	counterfeit currency	counterfeit currency	gold bars
At a crime scene, there is a closed crate containing gold bars, labeled as 'gold bars'. An investigator thinks it contains counterfeit currency. He opens the crate, and looks inside. The investigator closes the crate.	gold bars	counterfeit currency	gold bars
At a crime scene, there is a closed crate containing gold bars, labeled as 'gold bars'. An investigator thinks it contains counterfeit currency. Despite trying, he cannot open the crate. So, the investigator stops trying.	gold bars	counterfeit currency	gold bars

Stimuli	Correct answer	Option 1	Option 2
Mark has labeled a transparent jar, filled with honey, as 'peanut butter', as a prank. His brother finds the jar and reads its label. Despite trying, his brother cannot open the jar and stops trying.	honey	honey	peanut butter
Mark has labeled a non-transparent jar, filled with honey, as 'peanut butter', as a prank. His brother finds the jar and reads its label. Despite trying, his brother cannot open the jar and stops trying.	peanut butter	honey	peanut butter
Mark has labeled a transparent jar, filled with honey, as 'peanut butter', as a prank. His brother finds the jar and reads its label. Despite trying, his brother cannot open the jar and stops trying.	honey	honey	peanut butter
Tom has mistakenly labeled a non-transparent jar, containing sugar, as 'jelly'. Long after, Tom finds the jar, but he cannot remember what is in it. He reads the label carefully. So, he takes the jar.	jelly	jelly	sugar
Tom has mistakenly labeled a non-transparent jar, containing sugar, as 'jelly'. Long after, Tom finds the jar, but he cannot remember what is in it. He reads the label and opens it. He closes it.	sugar	jelly	sugar
Tom has mistakenly labeled a non-transparent jar, containing sugar, as 'jelly'. Long after, Tom finds the jar, but he cannot remember what is in it. He reads the label carefully. So, he takes the jar.	sugar	jelly	sugar
Emma wants to surprise her sister, Lisa, with a gift with toys, labeled 'candy'. Their parents spoil the surprise. Lisa finds the gift, reads the label, and opens the gift. So, Lisa closes the gift.	toys	toys	candy
Emma wants to surprise her sister, Lisa, with a gift with toys, labeled 'candy'. Their parents also take part. Lisa finds the gift, reads its label, but she cannot open it. So, she stops trying.	candy	toys	candy

Stimuli	Correct answer	Option 1	Option 2
Emma wants to surprise her sister, Lisa, with a gift with toys, labeled 'candy'. Their parents spoil the surprise. Lisa finds the gift, reads the label, and opens the gift. So, Lisa closes the gift.	toys	toys	candy
John washes his hands with soap from a bottle, labeled 'rose soap', that contains rosemary soap. Having a stuffy nose, he has lost his sense of smell. He reads the label. He exits the room.	rose soap	rose soap	rosemary soap
John washes his hands with soap from a bottle, labeled 'rose soap', that actually contains rosemary soap. Noticing a rosemary smell, he takes the bottle, and carefully reads its ingredients. He puts the bottle down.	rosemary soap	rose soap	rosemary soap
John washes his hands with soap from a bottle, labeled 'rose soap', that contains rosemary soap. Having a stuffy nose, he has lost his sense of smell. He reads the label. He exits the room.	rosemary soap	rose soap	rosemary soap
On a desk, there is a USB drive with text documents, labeled 'text documents'. A new intern finds the drive. Seeing it for the first time, he reads the label. So, he takes the drive.	text documents	text documents	audio documents
On a desk, there is a USB drive with text documents, labeled 'audio documents'. A new intern finds the drive. Seeing it for the first time, he reads the label. So, he takes the drive.	audio documents	text documents	audio documents
On a desk, there is a USB drive with text documents, labeled 'text documents'. A new intern finds the drive. Seeing it for the first time, he reads the label. So, he takes the drive.	text documents	text documents	audio documents
In the kitchen, there is a non-transparent jar with pickles. Its label says 'olives', in French. Lisa notices it, reading the label which is clearly written in her native language. So, she takes the jar.	olives	olives	pickles

Stimuli	Correct answer	Option 1	Option 2
In the kitchen, there is a transparent jar with pickles. Its label says 'olives', in French. Lisa notices it, trying to read the label, but she does not speak French. So, she takes the jar.	pickles	olives	pickles
In the kitchen, there is a non-transparent jar, full of pickles, with no olives. Its label says 'olives' in French and not 'pickles'. Lisa walks in and notices the jar. She has never seen the jar before. She reads the label which is written clearly in her native language, but she cannot open the jar.	pickles	olives	pickles
Michael ordered headphones online. The distribution center ships a box of chargers in a transparent bag, labeled 'headphones', at his front door. He eventually finds the package, reading its label. He takes the bag inside.	chargers	chargers	headphones
Michael ordered headphones online. The distribution center ships a box of chargers in a non-transparent bag, labeled 'headphones', at his front door. He eventually finds the package, reading its label. He takes the bag inside.	headphones	chargers	headphones
Michael ordered headphones online. The distribution center ships a box of chargers in a transparent bag, labeled 'headphones', at his front door. He eventually finds the package, reading its label. He takes the bag inside.	chargers	chargers	headphones
Tom hands a gift containing a hat to Lucy, saying it contains gloves. Their aunt, whom Lucy trusts, says 'That is actually a scarf'. Lucy cannot open the gift. She gives up and stops trying.	scarf	scarf	hats
Tom hands a gift containing a hat to Lucy, saying it contains gloves. Their aunt, who bought the gift, says 'That is actually a hat'. Lucy looks inside the gift. So, she closes the gift.	hats	gloves	hats

Stimuli	Correct answer	Option 1	Option 2
Tom hands a gift containing a hat to Lucy, saying it contains gloves. Their aunt, whom Lucy trusts, says 'That is actually a scarf'. Lucy cannot open the gift. She gives up and stops trying.	hats	scarf	hats
A transparent plastic bin, containing electronics, arrives at a shipping warehouse. The inventory says the bin contains books. Jane, an employee, has to ship the bin, and she sees it. So, she takes the bin.	electronics	electronics	books
A non-transparent plastic bin, containing electronics, arrives at a shipping warehouse. The inventory says the bin contains books. Jane, an employee, has to ship the bin, and she sees it. So, she takes the bin.	books	electronics	books
A transparent plastic bin, containing electronics, arrives at a shipping warehouse. The inventory says the bin contains books. Jane, an employee, has to ship the bin, and she sees it. So, she takes the bin.	electronics	electronics	books
On a desk, there is a non-transparent envelope labeled 'staples', filled with paper clips. An intern reads the label, and hears a trusted co-worker saying 'That envelope contains many pens'. The intern takes the envelope.	pens	pens	paper clips
On a desk, there is a transparent envelope labeled 'staples', filled with paper clips. An intern reads the label, and hears a co-worker jokingly saying 'That envelope contains many pens'. The intern takes the envelope.	paper clips	pens	paper clips
On a desk, there is a non-transparent envelope labeled 'staples', filled with paper clips. An intern reads the label, and hears a trusted co-worker saying 'That envelope contains many pens'. The intern takes the envelope.	paper clips	pens	paper clips

Stimuli	Correct answer	Option 1	Option 2
In his father's garage, Kevin finds CDs labeled 'blues', that are actually of classical music. He reads the label. His father says 'Those contain classical music'. Kevin trusts his father. So, Kevin takes the CDs.	classical music	classical music	blues music
In his father's garage, Kevin finds CDs labeled 'blues', that actually contain classical music. He reads the label. His father says 'I forgot they existed'. Kevin reads the label again. So, Kevin takes the CDs.	blues music	classical music	blues music
In his father's garage, Kevin finds CDs labeled 'blues', that are actually of classical music. He reads the label. His father says 'Those contain classical music'. Kevin trusts his father. So, Kevin takes the CDs.	classical music	classical music	blues music
In a museum, there is a non-transparent black box labeled 'ancient coins', that actually contains old maps. A curator reads the label and tries to open the box, but fails. So, the curator stops trying.	ancient coins	ancient coins	old maps
In a museum, there is a non-transparent black box labeled 'ancient coins', that actually contains old maps. A curator reads the label and opens the box, looking inside of it. The curator closes the box.	old maps	ancient coins	old maps
In a museum, there is a non-transparent black box labeled 'ancient coins', that actually contains old maps. A curator reads the label and tries to open the box, but fails. So, the curator stops trying.	old maps	ancient coins	old maps
At a library, there is a book with horror stories, titled 'Fairy Tales'. The librarian, a fairy tale enthusiast, is shelving books. She reads the title, and keeps the book. She takes it with her.	fairy tales	fairy tales	horror stories
At a library, there is a book with horror stories, titled 'Fairy Tales'. The librarian, a fairy tale enthusiast, is shelving books. She reads the title, and keeps the book. She takes it with her.	fairy tales	fairy tales	horror stories

Stimuli	Correct answer	Option 1	Option 2
At a library, there is a book with horror stories, titled 'Fairy Tales'. The librarian, a fairy tale enthusiast, is shelving books. She reads the title, and keeps the book. She takes it with her.	horror stories	fairy tales	horror stories
In a warehouse, there is a crate that contains metal tools, labeled 'fragile glassware'. An artist, in need of delicate items, carefully moves the crate to a designated work area, before taking the crate home.	fragile glassware	fragile glassware	metal tools
In a warehouse, there is a crate that contains metal tools, labeled 'fragile glassware'. An artist, in need of delicate items, carefully moves the crate to a designated work area, before taking the crate home.	fragile glassware	fragile glassware	metal tools
In a warehouse, there is a crate that contains metal tools, labeled 'fragile glassware'. An artist, in need of delicate items, carefully moves the crate to a designated work area, before taking the crate home.	metal tools	fragile glassware	metal tools
In a park, a child sees someone looking very sad. The child, wishing to cheer the person up, approaches with a flower and offers it to them. The person smiles, feeling happier and less lonely.	to brighten the person's day	to brighten the person's day	to make the person smile
In a park, a child sees someone looking very sad. The child, wishing to cheer the person up, approaches with a flower and offers it to them. The person smiles, feeling happier and less lonely.	to brighten the person's day	to brighten the person's day	to make the person smile
In a park, a child sees someone looking very sad. The child, wishing to cheer the person up, approaches with a flower and offers it to them. The person smiles, feeling happier and less lonely.	brightened the person's day	brightened the person's day	made the person smile

Stimuli	Correct answer	Option 1	Option 2
In an office, an employee looks stressed and overwhelmed with work. A colleague offers to assist with their tasks. The employee feels relieved, and directly expresses gratitude for the support. The employee accepts the help.	to lighten the colleague's workload	to lighten the colleague's workload	to ease the colleague's stress
In an office, an employee looks stressed and overwhelmed with work. A colleague offers to assist with their tasks. The employee feels relieved, and directly expresses gratitude for the support. The employee accepts the help.	to lighten the colleague's workload	to lighten the colleague's workload	to ease the colleague's stress
In an office, an employee looks stressed and overwhelmed with work. A colleague offers to assist with their tasks. The employee feels relieved, and directly expresses gratitude for the support. The employee accepts the help.	lightened the colleague's workload	lightened the colleague's workload	eased the colleague's stress
In a classroom, there is a big book labeled 'Math Facts' that actually contains historical dates about Mathematics. A student, who loves math formulae, reads its title and approaches it. He takes the book home.	math formulae	historical dates	math formulae
In a classroom, there is a big book labeled 'Math Facts' that actually contains historical dates about Mathematics. A student, who loves math formulae, reads its title and approaches it. He takes the book home.	math formulae	historical dates	math formulae
In a classroom, there is a big book labeled 'Math Facts' that actually contains historical dates about Mathematics. A student, who loves math formulae, reads its title and approaches it. He takes the book home.	historical dates	historical dates	math formulae
In a doctor's office, there is a jar with vitamins, labeled 'Candies'. A child, who loves candies, reads the label. The child wants to open the jar for its contents. He waits to do so.	candies	vitamins	candies

Stimuli	Correct answer	Option 1	Option 2
In a doctor's office, there is a jar with vitamins, labeled 'Candies'. A child, who loves candies, reads the label. The child wants to open the jar for its contents. He waits to do so.	candies	vitamins	candies
In a doctor's office, there is a jar with vitamins, labeled 'Candies'. A child, who loves candies, reads the label. The child wants to open the jar for its contents. He waits to do so.	vitamins	vitamins	candies
At a library, there is a book with horror stories, titled 'Fairy Tales'. The librarian, a fairy tale enthusiast, is shelving books. She reads the title, and keeps the book. She takes it with her.	fairy tales	fairy tales	horror stories
At a library, there is a book with horror stories, titled 'Fairy Tales'. The librarian, a fairy tale enthusiast, is shelving books. She reads the title, and keeps the book. She takes it with her.	fairy tales	fairy tales	horror stories
At a library, there is a book with horror stories, titled 'Fairy Tales'. The librarian, a fairy tale enthusiast, is shelving books. She reads the title, and keeps the book. She takes it with her.	horror stories	fairy tales	horror stories
In a warehouse, there is a crate that contains metal tools, labeled 'fragile glassware'. An artist, in need of delicate items, carefully moves the crate to a designated work area, before taking the crate home.	fragile glassware	fragile glassware	metal tools
In a warehouse, there is a crate that contains metal tools, labeled 'fragile glassware'. An artist, in need of delicate items, carefully moves the crate to a designated work area, before taking the crate home.	fragile glassware	fragile glassware	metal tools
In a warehouse, there is a crate that contains metal tools, labeled 'fragile glassware'. An artist, in need of delicate items, carefully moves the crate to a designated work area, before taking the crate home.	metal tools	fragile glassware	metal tools

Stimuli	Correct answer	Option 1	Option 2
In a park, a child sees someone looking very sad. The child, wishing to cheer the person up, approaches with a flower and offers it to them. The person smiles, feeling happier and less lonely.	to brighten the person's day	to brighten the person's day	to make the person smile
In a park, a child sees someone looking very sad. The child, wishing to cheer the person up, approaches with a flower and offers it to them. The person smiles, feeling happier and less lonely.	to brighten the person's day	to brighten the person's day	to make the person smile
In a park, a child sees someone looking very sad. The child, wishing to cheer the person up, approaches with a flower and offers it to them. The person smiles, feeling happier and less lonely.	brightened the person's day	brightened the person's day	made the person smile
In an office, an employee looks stressed and overwhelmed with work. A colleague offers to assist with their tasks. The employee feels relieved, and directly expresses gratitude for the support. The employee accepts the help.	to lighten the colleague's workload	to lighten the colleague's workload	to ease the colleague's stress
In an office, an employee looks stressed and overwhelmed with work. A colleague offers to assist with their tasks. The employee feels relieved, and directly expresses gratitude for the support. The employee accepts the help.	to lighten the colleague's workload	to lighten the colleague's workload	to ease the colleague's stress
In an office, an employee looks stressed and overwhelmed with work. A colleague offers to assist with their tasks. The employee feels relieved, and directly expresses gratitude for the support. The employee accepts the help.	lightened the colleague's workload	lightened the colleague's workload	eased the colleague's stress
In a classroom, there is a big book labeled 'Math Facts' that actually contains historical dates about Mathematics. A student, who loves math formulae, reads its title and approaches it. He takes the book home.	math formulae	historical dates	math formulae

Stimuli	Correct answer	Option 1	Option 2
In a classroom, there is a big book labeled 'Math Facts' that actually contains historical dates about Mathematics. A student, who loves math formulae, reads its title and approaches it. He takes the book home.	math formulae	historical dates	math formulae
In a classroom, there is a big book labeled 'Math Facts' that actually contains historical dates about Mathematics. A student, who loves math formulae, reads its title and approaches it. He takes the book home.	historical dates	historical dates	math formulae
In a doctor's office, there is a jar with vitamins, labeled 'Candies'. A child, who loves candies, reads the label. The child wants to open the jar for its contents. He waits to do so.	candies	vitamins	candies
In a doctor's office, there is a jar with vitamins, labeled 'Candies'. A child, who loves candies, reads the label. The child wants to open the jar for its contents. He waits to do so.	candies	vitamins	candies
In a doctor's office, there is a jar with vitamins, labeled 'Candies'. A child, who loves candies, reads the label. The child wants to open the jar for its contents. He waits to do so.	vitamins	vitamins	candies
At a library, there is a book with horror stories, titled 'Fairy Tales'. The librarian, a fairy tale enthusiast, is shelving books. She reads the title, and keeps the book. She takes it with her.	fairy tales	fairy tales	horror stories
At a library, there is a book with horror stories, titled 'Fairy Tales'. The librarian, a fairy tale enthusiast, is shelving books. She reads the title, and keeps the book. She takes it with her.	fairy tales	fairy tales	horror stories

Stimuli	Correct answer	Option 1	Option 2
At a library, there is a book with horror stories, titled 'Fairy Tales'. The librarian, a fairy tale enthusiast, is shelving books. She reads the title, and keeps the book. She takes it with her.	horror stories	fairy tales	horror stories
In a warehouse, there is a crate that contains metal tools, labeled 'fragile glassware'. An artist, in need of delicate items, carefully moves the crate to a designated work area, before taking the crate home.	fragile glassware	fragile glassware	metal tools
In a warehouse, there is a crate that contains metal tools, labeled 'fragile glassware'. An artist, in need of delicate items, carefully moves the crate to a designated work area, before taking the crate home.	fragile glassware	fragile glassware	metal tools
In a warehouse, there is a crate that contains metal tools, labeled 'fragile glassware'. An artist, in need of delicate items, carefully moves the crate to a designated work area, before taking the crate home.	metal tools	fragile glassware	metal tools
In a park, a child sees someone looking very sad. The child, wishing to cheer the person up, approaches with a flower and offers it to them. The person smiles, feeling happier and less lonely.	to brighten the person's day	to brighten the person's day	to make the person smile
In a park, a child sees someone looking very sad. The child, wishing to cheer the person up, approaches with a flower and offers it to them. The person smiles, feeling happier and less lonely.	to brighten the person's day	to brighten the person's day	to make the person smile
In a park, a child sees someone looking very sad. The child, wishing to cheer the person up, approaches with a flower and offers it to them. The person smiles, feeling happier and less lonely.	brightened the person's day	brightened the person's day	made the person smile

Stimuli	Correct answer	Option 1	Option 2
In an office, an employee looks stressed and overwhelmed with work. A colleague offers to assist with their tasks. The employee feels relieved, and directly expresses gratitude for the support. The employee accepts the help.	to lighten the colleague's workload	to lighten the colleague's workload	to ease the colleague's stress
In an office, an employee looks stressed and overwhelmed with work. A colleague offers to assist with their tasks. The employee feels relieved, and directly expresses gratitude for the support. The employee accepts the help.	to lighten the colleague's workload	to lighten the colleague's workload	to ease the colleague's stress
In an office, an employee looks stressed and overwhelmed with work. A colleague offers to assist with their tasks. The employee feels relieved, and directly expresses gratitude for the support. The employee accepts the help.	lightened the colleague's workload	lightened the colleague's workload	eased the colleague's stress
In a classroom, there is a big book labeled 'Math Facts' that actually contains historical dates about Mathematics. A student, who loves math formulae, reads its title and approaches it. He takes the book home.	math formulae	historical dates	math formulae
In a classroom, there is a big book labeled 'Math Facts' that actually contains historical dates about Mathematics. A student, who loves math formulae, reads its title and approaches it. He takes the book home.	math formulae	historical dates	math formulae
In a classroom, there is a big book labeled 'Math Facts' that actually contains historical dates about Mathematics. A student, who loves math formulae, reads its title and approaches it. He takes the book home.	historical dates	historical dates	math formulae
In a doctor's office, there is a jar with vitamins, labeled 'Candies'. A child, who loves candies, reads the label. The child wants to open the jar for its contents. He waits to do so.	candies	vitamins	candies

Stimuli	Correct answer	Option 1	Option 2
In a doctor's office, there is a jar with vitamins, labeled 'Candies'. A child, who loves candies, reads the label. The child wants to open the jar for its contents. He waits to do so.	candies	vitamins	candies
In a doctor's office, there is a jar with vitamins, labeled 'Candies'. A child, who loves candies, reads the label. The child wants to open the jar for its contents. He waits to do so.	vitamins	vitamins	candies

Table A.1: Table representing the stimuli created for the FBT task, their correct answers, and the options participants had to choose from. The stimuli could be reviewed under Column *Stimuli*, while their correct answers are specified under Column *Correct answer*. The first and second options to choose from are shown in Columns *Option 1* and *Option 2*.

Bibliography

[1] Natalie Shapira, Mosh Levy, Seyed Hossein Alavi, Xuhui Zhou, Yejin Choi, Yoav Goldberg, Maarten Sap, and Vered Shwartz. Clever hans or neural theory of mind? stress testing social reasoning in large language models. In Yvette Graham and Matthew Purver, editors, *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2257–2273, St. Julian's, Malta, March 2024. Association for Computational Linguistics. URL https://aclanthology.org/2024.eacl-long.138.

[2] Alan Agresti. *Categorical data analysis*, volume 792. John Wiley & Sons, 2012.

[3] Alan M Leslie, Ori Friedman, and Tim P German. Core mechanisms in 'theory of mind'. *Trends in cognitive sciences*, 8(12):528–533, 2004.

[4] Henry M Wellman, David Cross, and Julanne Watson. Meta-analysis of theory-of-mind development: The truth about false belief. *Child development*, 72(3):655–684, 2001.

[5] Jill De Villiers. The interface of language and theory of mind. *Lingua*, 117(11):1858–1878, 2007.

[6] Nikolos Gurney, David V. Pynadath, and Volkan Ustun. Spontaneous theory of mind for artificial intelligence, 2024.

[7] John Hewitt and Michael Cohen. Exploring roberta's theory of mind through textual entailment. 2021. URL https://api.semanticscholar.org/CorpusID:235357901.

[8] Max J. van Duijn, Bram M. A. van Dijk, Tom Kouwenhoven, Werner de Valk, Marco R. Spruit, and Peter van der Putten. Theory of mind in large language models: Examining performance of 11 state-of-the-art models vs. children aged 7-10 on advanced tests, 2023. URL https://arxiv.org/abs/2310.20320.

[9] Tomer Ullman. Large language models fail on trivial alterations to theory-of-mind tasks, 2023.

[10] Melanie Sclar, Sachin Kumar, Peter West, Alane Suhr, Yejin Choi, and Yulia Tsvetkov. Minding language models' (lack of) theory of mind: A plug-and-play multi-character belief tracker, 2023.

[11] Maarten Sap, Ronan LeBras, Daniel Fried, and Yejin Choi. Neural theory-of-mind? on the limits of social intelligence in large lms, 2023.

[12] Bart Holterman and Kees van Deemter. Does chatgpt have theory of mind?, 2023.

[13] Mudit Verma, Siddhant Bhambri, and Subbarao Kambhampati. Theory of mind abilities of large language models in human-robot interaction: An illusion? In *Companion of the 2024 ACM/IEEE International Conference on Human-Robot Interaction*, pages 36–45, 2024.

[14] Henry M Wellman. Developing a theory of mind. *The Wiley-Blackwell handbook of childhood cognitive development*, 2:258–284, 2011.

[15] Sean Trott, Cameron J. Jones, Tyler A. Chang, James A. Michaelov, and Benjamin K. Bergen. Do large language models know what humans know? *Cognitive science*, 47 7: e13309, 2022. URL https://api.semanticscholar.org/CorpusID:252089182.

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

[17] Karen Milligan, Janet Wilde Astington, and Lisa Ain Dack. Language and theory of mind: Meta-analysis of the relation between language ability and false-belief understanding. *Child development*, 78(2):622–646, 2007.

[18] Courtney Melinda Hale and Helen Tager-Flusberg. The influence of language on theory of mind: A training study. *Developmental science*, 6(3):346–359, 2003.

[19] Stephanie Durrleman, Anamaria Bentea, Andreea Prisecaru, Evelyne Thommen, and Hélène Delage. Training syntax to enhance theory of mind in children with asd. *Journal of Autism and Developmental Disorders*, 53(6):2444–2457, 2023.

[20] M Jeffrey Farrar, Joann P Benigno, Virginia Tompkins, and Nicholas A Gage. Are there different pathways to explicit false belief understanding? general language and complementation in typical and atypical children. *Cognitive Development*, 43:49–66, 2017.

[21] Sanne JM Kuijper, Catharina A Hartman, and Petra Hendriks. Children's pronoun interpretation problems are related to theory of mind and inhibition, but not working memory. *Frontiers in psychology*, 12:610401, 2021.

[22] Petra Hendriks and Jennifer Spenader. When production precedes comprehension: An optimization approach to the acquisition of pronouns. *Language Acquisition*, 13:319 – 348, 2006. URL https://api.semanticscholar.org/CorpusID:15563213.

[23] Robert M. Gordon. Folk psychology as simulation. *Mind & Language*, 1:158–171, 1986. URL https://api.semanticscholar.org/CorpusID:144759272.

[24] Mohsen Jamali, Ziv M. Williams, and Jing Cai. Unveiling theory of mind in large language models: A parallel to single neurons in the human brain, 2023.

[25] Winnie Street. Llm theory of mind and alignment: Opportunities and risks, 2024. URL https://arxiv.org/abs/2405.08154.

[26] Francesca Bianco and Dimitri Ognibene. Robot learning theory of mind through self-observation: Exploiting the intentions-beliefs synergy, 2022.

[27] Kanishk Gandhi, Jan-Philipp Fränken, Tobias Gerstenberg, and Noah D. Goodman. Understanding social reasoning in language models with language models, 2023.

[28] Jessica Williams, Stephen M. Fiore, and Florian Jentsch. Supporting artificial social intelligence with theory of mind. *Frontiers in Artificial Intelligence*, 5, 2022. ISSN 2624-8212. doi: 10.3389/frai.2022.750763. URL https://www.frontiersin.org/articles/10.3389/frai.2022.750763.

[29] Michal Kosinski. Evaluating large language models in theory of mind tasks, 2024.

[30] Pei Zhou, Aman Madaan, Srividya Pranavi Potharaju, Aditya Gupta, Kevin R. McKee, Ari Holtzman, Jay Pujara, Xiang Ren, Swaroop Mishra, Aida Nematzadeh, Shyam Upadhyay, and Manaal Faruqui. How far are large language models from agents with theory-of-mind? *ArXiv*, abs/2310.03051, 2023. URL https://api.semanticscholar.org/CorpusID:263671666.

[31] Neil C. Rabinowitz, Frank Perbet, H. Francis Song, Chiyuan Zhang, S. M. Ali Eslami, and Matthew Botvinick. Machine theory of mind, 2018.

[32] Wentao Zhu, Zhining Zhang, and Yizhou Wang. Language models represent beliefs of self and others, 2024.

[33] Guiyang Hou, Wenqi Zhang, Yongliang Shen, Linjuan Wu, and Weiming Lu. Timetom: Temporal space is the key to unlocking the door of large language models' theory-of-mind, 2024. URL https://arxiv.org/abs/2407.01455.

[34] Weizhi Tang and Vaishak Belle. Tom-lm: Delegating theory of mind reasoning to external symbolic executors in large language models, 2024. URL https://arxiv.org/abs/2404.15515.

[35] Pei Zhou, Aman Madaan, Srividya Pranavi Potharaju, Aditya Gupta, Kevin R. McKee, Ari Holtzman, Jay Pujara, Xiang Ren, Swaroop Mishra, Aida Nematzadeh, Shyam Upadhyay, and Manaal Faruqui. How far are large language models from agents with theory-of-mind?, 2023. URL https://arxiv.org/abs/2310.03051.

[36] Shima Rahimi Moghaddam and Christopher J. Honey. Boosting theory-of-mind performance in large language models via prompting, 2023. URL https://arxiv.org/abs/2304.11490.

[37] Maryam Amirizaniani, Elias Martin, Maryna Sivachenko, Afra Mashhadi, and Chirag Shah. Do llms exhibit human-like reasoning? evaluating theory of mind in llms for open-ended responses, 2024. URL https://arxiv.org/abs/2406.05659.

[38] Nunzio Lore, Alireza Sepehr Ilami, and Babak Heydari. Large model strategic thinking, small model efficiency: Transferring theory of mind in large language models, 2024. URL https://arxiv.org/abs/2408.05241.

[39] Khai Loong Aw, Syrielle Montariol, Badr AlKhamissi, Martin Schrimpf, and Antoine Bosselut. Instruction-tuning aligns llms to the human brain, 2023. URL https://arxiv.org/abs/2312.00575.

[40] Jennifer Hu, Kyle Mahowald, Gary Lupyan, Anna Ivanova, and Roger Levy. Language models align with human judgments on key grammatical constructions, 2024. URL https://arxiv.org/abs/2402.01676.

[41] Asaf Yehudai, Taelin Karidi, Gabriel Stanovsky, Ariel Goldstein, and Omri Abend. A nurse is blue and elephant is rugby: Cross domain alignment in large language models reveal human-like patterns, 2024. URL https://arxiv.org/abs/2405.14863.

[42] Cansu Sen, Thomas Hartvigsen, Biao Yin, Xiangnan Kong, and Elke Rundensteiner. Human attention maps for text classification: Do humans and neural networks focus on the same words? In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault, editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4596–4608, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.419. URL https://aclanthology.org/2020.acl-main.419.

[43] Xuesong Chen, Jiaxin Mao, Yiqun Liu, M. Zhang, and Shaoping Ma. Investigating human reading behavior during sentiment judgment. *International Journal of Machine Learning and Cybernetics*, 13:2283 – 2296, 2022. URL https://api.semanticscholar.org/CorpusID:247288120.

[44] Jeevithashree Divya Venkatesh, Aparajita Jaiswal, and Gaurav Nanda. Comparing human text classification performance and explainability with large language and machine learning models using eye-tracking. *Scientific Reports*, 14(1):14295, 2024.

[45] Elize Herrewijnen, Dong Nguyen, Floris Bex, and Kees van Deemter. Human-annotated rationales and explainable text classification: a survey. *Frontiers in Artificial Intelligence*, 7:1260952, 2024.

[46] Rachel Ostrand and Sara E Berger. Humans linguistically align to their conversational partners, and language models should too. In *ICML 2024 Workshop on LLMs and Cognition*, 2024. URL https://openreview.net/forum?id=JtgxazN6TM.

[47] Tatsuki Kuribayashi, Yohei Oseki, and Timothy Baldwin. Psychometric predictive power of large language models, 2024. URL https://arxiv.org/abs/2311.07484.

[48] Miriam Rivero-Contreras, David Saldaña, and Martina Micai. An introduction to theory of mind: Fundamental concepts and issues. In *The Theory of Mind Under Scrutiny: Psychopathology, Neuroscience, Philosophy of Mind and Artificial Intelligence*, pages 11–33. Springer, 2024.

[49] Cecilia Heyes. Submentalizing: I am not really reading your mind. *Perspectives on Psychological Science*, 9:131 – 143, 2014. URL https://api.semanticscholar.org/CorpusID:206778161.

[50] David Liu, Henry M Wellman, Twila Tardif, and Mark A Sabbagh. Theory of mind development in chinese children: a meta-analysis of false-belief understanding across cultures and languages. *Developmental psychology*, 44(2):523, 2008.

[51] Francesca Happé. Theory of mind and the self. *Annals of the New York Academy of Sciences*, 1001(1):134–144, 2003.

[52] Jodie A. Baird and Janet Wilde Astington. 23Introduction: Why Language Matters. In *Why Language Matters for Theory of Mind*. Oxford University Press, 04 2005. ISBN 9780195159912. doi: 10.1093/acprof:oso/9780195159912.003.0001. URL https://doi.org/10.1093/acprof:oso/9780195159912.003.0001.

[53] Jill G De Villiers and Peter A de Villiers. The role of language in theory of mind development. *Topics in Language Disorders*, 34(4):313–328, 2014.

[54] Katherine G. Nelson. Language pathways into the community of minds. 2005. URL https://api.semanticscholar.org/CorpusID:148547250.

[55] Janet Wilde Astington and Jodie A. Baird. 163Representational Development and False-Belief Understanding. In *Why Language Matters for Theory of Mind*. Oxford University Press, 04 2005. ISBN 9780195159912. doi: 10.1093/acprof:oso/9780195159912.003.0009. URL https://doi.org/10.1093/acprof:oso/9780195159912.003.0009.

[56] Brian W Miller. Using reading times and eye-movements to measure cognitive engagement. *Educational psychologist*, 50(1):31–42, 2015.

[57] Alain Morin, Eaman El-Sayed, and Famira Racy. Self-awareness, inner speech, and theory of mind in typical and asd individuals: A critical review. *Theory of mind: development in children, brain mechanisms and social implications. Nova Science Pub*, 2015.

[58] Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg, Harsha Nori, Hamid Palangi, Marco Tulio Ribeiro, and Yi Zhang. Sparks of artificial general intelligence: Early experiments with gpt-4, 2023.

[59] Matthew Le, Y-Lan Boureau, and Maximilian Nickel. Revisiting the evaluation of theory of mind through question answering. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5872–5877, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1598. URL https://www.aclweb.org/anthology/D19-1598.

[60] Maarten Sap, Hannah Rashkin, Derek Chen, Ronan Le Bras, and Yejin Choi. Social IQa: Commonsense reasoning about social interactions. In Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan, editors, *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4463–4473, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1454. URL https://aclanthology.org/D19-1454.

[61] Ilia Sucholutsky and Thomas L Griffiths. Why should we care if machines learn human-like representations?

[62] Hua Shen, Tiffany Knearem, Reshmi Ghosh, Kenan Alkiek, Kundan Krishna, Yachuan Liu, Ziqiao Ma, Savvas Petridis, Yi-Hao Peng, Li Qiwei, Sushrita Rakshit, Chenglei Si, Yutong Xie, Jeffrey P. Bigham, Frank Bentley, Joyce Chai, Zachary Lipton, Qiaozhu Mei, Rada Mihalcea, Michael Terry, Diyi Yang, Meredith Ringel Morris, Paul Resnick, and David Jurgens. Towards bidirectional human-ai alignment: A systematic review for clarifications, framework, and future directions, 2024. URL https://arxiv.org/abs/2406.09264.

[63] Ilia Sucholutsky and Tom Griffiths. Alignment with human representations supports robust few-shot learning. *Advances in Neural Information Processing Systems*, 36, 2024.

[64] Zhiyu Tan, Xiaomeng Yang, Luozheng Qin, Mengping Yang, Cheng Zhang, and Hao Li. Evalalign: Supervised fine-tuning multimodal llms with human-aligned data for evaluating text-to-image models, 2024. URL https://arxiv.org/abs/2406.16562.

[65] Zhilin Wang, Yu Ying Chiu, and Yu Cheung Chiu. Humanoid agents: Platform for simulating human-like generative agents, 2023. URL https://arxiv.org/abs/2310.05418.

[66] Lei Zhao, Yingyi Zhang, and Chengzhi Zhang. Does attention mechanism possess the feature of human reading? a perspective of sentiment classification task. *Aslib Journal of Information Management*, 75(1):20–43, May 2022. ISSN 2050-3806. doi: 10.1108/ajim-12-2021-0385. URL http://dx.doi.org/10.1108/AJIM-12-2021-0385.

[67] Dongyu Zhang, Cansu Sen, Jidapa Thadajarassiri, Thomas Hartvigsen, Xiangnan Kong, and Elke Rundensteiner. Human-like explanation for text classification with limited attention supervision. In *2021 ieee international conference on big data (big data)*, pages 957–967. IEEE, 2021.

[68] Katherine M. Collins, Catherine Wong, Jiahai Feng, Megan Wei, and Joshua B. Tenenbaum. Structured, flexible, and robust: benchmarking and improving large language models towards more human-like behavior in out-of-distribution reasoning tasks, 2022. URL https://arxiv.org/abs/2205.05718.

[69] Haoyan Luo and Lucia Specia. From understanding to utilization: A survey on explainability for large language models, 2024. URL https://arxiv.org/abs/2401.12874.

[70] Haiyan Zhao, Hanjie Chen, Fan Yang, Ninghao Liu, Huiqi Deng, Hengyi Cai, Shuaiqiang Wang, Dawei Yin, and Mengnan Du. Explainability for large language models: A survey. *ACM Transactions on Intelligent Systems and Technology*, 15(2):1–38, 2024.

[71] Manas Gaur and Amit Sheth. Building trustworthy neurosymbolic ai systems: Consistency, reliability, explainability, and safety. *AI Magazine*, 45(1):139–155, 2024.

[72] Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. In Doina Precup and Yee Whye Teh, editors, *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 3319–3328. PMLR, 06–11 Aug 2017. URL https://proceedings.mlr.press/v70/sundararajan17a.html.

[73] Ian E. Nielsen, Dimah Dera, Ghulam Rasool, Ravi P. Ramachandran, and Nidhal Carla Bouaynaya. Robust explainability: A tutorial on gradient-based attribution methods for deep neural networks. *IEEE Signal Processing Magazine*, 39(4):73–84, July 2022. ISSN 1558-0792. doi: 10.1109/msp.2022.3142719. URL http://dx.doi.org/10.1109/MSP.2022.3142719.

[74] Xuansheng Wu, Haiyan Zhao, Yaochen Zhu, Yucheng Shi, Fan Yang, Tianming Liu, Xiaoming Zhai, Wenlin Yao, Jundong Li, Mengnan Du, et al. Usable xai: 10 strategies towards exploiting explainability in the llm era. *arXiv preprint arXiv:2403.08946*, 2024.

[75] Marco Ancona, Enea Ceolini, Cengiz Öztireli, and Markus Gross. *Gradient-Based Attribution Methods*, pages 169–191. Springer International Publishing, Cham, 2019. ISBN 978-3-030-28954-6. doi: 10.1007/978-3-030-28954-6_9. URL https://doi.org/10.1007/978-3-030-28954-6_9.

[76] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. "why should i trust you?": Explaining the predictions of any classifier, 2016. URL https://arxiv.org/abs/1602.04938.

[77] Scott Lundberg and Su-In Lee. A unified approach to interpreting model predictions, 2017. URL https://arxiv.org/abs/1705.07874.

[78] Wojciech Samek, Thomas Wiegand, and Klaus-Robert Müller. Explainable artificial intelligence: Understanding, visualizing and interpreting deep learning models, 2017. URL https://arxiv.org/abs/1708.08296.

[79] Mohsen Fayyaz, Fan Yin, Jiao Sun, and Nanyun Peng. Evaluating human alignment and model faithfulness of llm rationale, 2024. URL https://arxiv.org/abs/2407.00219.

[80] EMMA MARSDEN, SOPHIE THOMPSON, and LUKE PLONSKY. A methodological synthesis of self-paced reading in second language research. *Applied Psycholinguistics*, 39 (5):861–904, 2018. doi: 10.1017/S0142716418000036.

[81] Jill Jegerski. *Self-paced reading*, pages 20–49. 01 2014.

[82] Marcel Adam Just, Patricia A. Carpenter, and Jacqueline D. Woolley. Paradigms and processes in reading comprehension. *Journal of Experimental Psychology: General*, 111: 228–238, 1982. URL https://api.semanticscholar.org/CorpusID:197658850.

[83] Else Havik, Leah Roberts, Roeland Van Hout, Robert Schreuder, and Marco Haverkort. Processing subject-object ambiguities in the l2: A self-paced reading study with german l2 learners of dutch. *Language Learning*, 59(1):73–112, 2009.

[84] Theodoros Marinis. Online methods in second language acquisition. *The encyclopedia of applied linguistics*, pages 4307–4315, 2013.

[85] Fernanda Ferreira and John M. Henderson. Use of verb information in syntactic parsing: evidence from eye movements and word-by-word self-paced reading. *Journal of experimental psychology. Learning, memory, and cognition*, 16 4:555–68, 1990. URL https://api.semanticscholar.org/CorpusID:17358216.

[86] Menno van der Schoot, Albert Reijntjes, and Ernest CDM van Lieshout. How do children deal with inconsistencies in text? an eye fixation and self-paced reading study in good and poor reading comprehenders. *Reading and Writing*, 25:1665–1690, 2012.

[87] Sean Trott, Cameron Jones, Tyler Chang, James Michaelov, and Benjamin Bergen. Do large language models know what humans know?, 2023.

[88] Michal Kosinski. Evaluating large language models in theory of mind tasks. *arXiv e-prints*, pages arXiv–2302, 2023.

[89] Sanne Kuijper, Catharina Hartman, and Petra Hendriks. Children's pronoun interpretation problems are related to theory of mind and inhibition, but not working memory. *Frontiers in Psychology*, 12:610401, 06 2021. doi: 10.3389/fpsyg.2021.610401.

[90] Alain Morin, Famira Racy, and Eaman El-Sayed. *Self-awareness, inner speech, and theory of mind in typical and ASD individuals: A critical review*, pages 43–113. 01 2015. ISBN 978-1-63463-856-2.

[91] Jeremy Zehr and Florian Schwarz. Penncontroller for internet based experiments (ibex), 2018. URL https://doi.org/10.17605/OSF.IO/MD832. DOI: 10.17605/OSF.IO/MD832.

[92] AI@Meta. Llama 3 model card. 2024. URL https://github.com/meta-llama/llama3/blob/main/MODEL_CARD.md.

[93] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692, 2019. URL http://arxiv.org/abs/1907.11692.

[94] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding, 2019. URL https://arxiv.org/abs/1810.04805.

[95] Ce Zhou, Qian Li, Chen Li, Jun Yu, Yixin Liu, Guangjing Wang, Kai Zhang, Cheng Ji, Qiben Yan, Lifang He, Hao Peng, Jianxin Li, Jia Wu, Ziwei Liu, Pengtao Xie, Caiming Xiong, Jian Pei, Philip S. Yu, and Lichao Sun. A comprehensive survey on pretrained foundation models: A history from bert to chatgpt, 2023. URL https://arxiv.org/abs/2302.09419.

[96] Pranav Rajpurkar, Robin Jia, and Percy Liang. Know what you don't know: Unanswerable questions for squad, 2018. URL https://arxiv.org/abs/1806.03822.

[97] Narine Kokhlikyan, Vivek Miglani, Miguel Martin, Edward Wang, Bilal Alsallakh, Jonathan Reynolds, Alexander Melnikov, Natalia Kliushkina, Carlos Araya, Siqi Yan, and Orion Reblitz-Richardson. Captum: A unified and generic model interpretability library for pytorch, 2020. URL https://arxiv.org/abs/2009.07896.

[98] Harsha Nori, Samuel Jenkins, Paul Koch, and Rich Caruana. Interpretml: A unified framework for machine learning interpretability. *arXiv preprint arXiv:1909.09223*, 2019.

[99] Catherine Thevenot, Michel Devidal, Pierre Barrouillet, and Michel Fayol. Why does placing the question before an arithmetic word problem improve performance? a situation model account. *Quarterly journal of experimental psychology*, 60(1):43–56, 2007.

[100] Laurice M Joseph, Sheila Alber-Morgan, Jennifer Cullen, and Christina Rouse. The effects of self-questioning on reading comprehension: A literature review. *Reading & Writing Quarterly*, 32(2):152–173, 2016.

[101] Ian A Apperly, Elisa Back, Dana Samson, and Lisa France. The cost of thinking about false beliefs: Evidence from adults' performance on a non-inferential theory of mind task. *Cognition*, 106(3):1093–1108, 2008.

[102] Margaret C. McKinnon and Morris Moscovitch. Domain-general contributions to social reasoning: Theory of mind and deontic reasoning re-explored. *Cognition*, 102(2):179–218, 2007. ISSN 0010-0277. doi: https://doi.org/10.1016/j.cognition.2005.12.011. URL https://www.sciencedirect.com/science/article/pii/S0010027706000035.

[103] Serena Lecce and Federica Bianco. Working memory predicts changes in children's theory of mind during middle childhood: A training study. *Cognitive Development*, 47:71–81, 2018.

[104] Fernanda Ferreira and Charles Clifton Jr. The independence of syntactic processing. *Journal of memory and language*, 25(3):348–368, 1986.

[105] Jill Jegerski. *Self-paced reading*, pages 20–49. 01 2014.

[106] Rajvardhan Patil and Venkat Gudivada. A review of current trends, techniques, and challenges in large language models (llms). *Applied Sciences*, 14(5):2074, 2024.

[107] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners, 2020. URL https://arxiv.org/abs/2005.14165.

[108] Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, Tom Hennigan, Eric Noland, Katie Millican, George van den Driessche, Bogdan Damoc, Aurelia Guy, Simon Osindero, Karen Simonyan, Erich Elsen, Jack W. Rae, Oriol Vinyals, and Laurent Sifre. Training compute-optimal large language models, 2022. URL https://arxiv.org/abs/2203.15556.

[109] Niklas Muennighoff, Alexander M. Rush, Boaz Barak, Teven Le Scao, Aleksandra Piktus, Nouamane Tazi, Sampo Pyysalo, Thomas Wolf, and Colin Raffel. Scaling data-constrained language models, 2023. URL https://arxiv.org/abs/2305.16264.

[110] Yichu Zhou and Vivek Srikumar. A closer look at how fine-tuning changes bert, 2022. URL https://arxiv.org/abs/2106.14282.

[111] Amil Merchant, Elahe Rahimtoroghi, Ellie Pavlick, and Ian Tenney. What happens to bert embeddings during fine-tuning?, 2020. URL https://arxiv.org/abs/2004.14448.

[112] Songyang Gao, Qiming Ge, Wei Shen, Shihan Dou, Junjie Ye, Xiao Wang, Rui Zheng, Yicheng Zou, Zhi Chen, Hang Yan, Qi Zhang, and Dahua Lin. Linear alignment: A closed-form solution for aligning human preferences without tuning and feedback, 2024. URL https://arxiv.org/abs/2401.11458.

[113] Yao Rong, Wenjia Xu, Zeynep Akata, and Enkelejda Kasneci. Human attention in fine-grained classification, 2021. URL https://arxiv.org/abs/2111.01628.

[114] Zhangyue Yin, Qiushi Sun, Qipeng Guo, Jiawen Wu, Xipeng Qiu, and Xuanjing Huang. Do large language models know what they don't know?, 2023.