Abstract

To many people, the ability to read another's mind is the ultimate superpower. The ability to see into how a person functions, why they express themselves the way they do, or perhaps what they think of you can be a very desirable trait. However, in everyday life, the closest one could get to this superpower is possessing a Theory of Mind (ToM), the capacity to ascribe mental states to others. ToM describes how humans infer and interpret actions, expressions, behaviours and acknowledge that others have different mental states. In this paper, our goal is to enhance Theory of Mind in Large Language Models (LLMs) by leveraging four datasets that pertain to ToM: emotional recognition, motivation recognition, false belief recognition, and epistemic logic. We aim to fine-tune the popular LLMs Falcon-7b and GPT-3.5 Turbo with these datasets.

1 Introduction

Theory of Mind (ToM) is crucial to the human experience, whether or not a person possesses ToM can determine if they understand how to socialize, communicate, and comprehend effectively when interacting with other people. Inferring and interpreting behaviours and expressions is another important aspect of ToM. This showcases an understanding that every person is their own entity capable of acting in a way that is different from one's own set of behavioural patterns. ToM is a complex skill developed over a number of years. It is both a naturally occurring and learned skill. A meta-study (Hofmann et al., 2016) showed that ToM can be improved in human subjects through training. Generally, ToM is not a skill that is learned through any direct means, suggesting that perhaps it is acquired through some amount of transfer learning.

ToM is a growing field of interest. Much of the currently available literature assesses whether or not Large Language Models (LLMs) possess any sort of indication of ToM capabilities. "...results show that models struggle substantially at (...) Theory of Mind tasks, with well-below-human accuracies…"(Sap et al., 2022), there have been many datasets developed with the intent to try to improve and test ToM in LLMs. Researchers in this area have focused on prompting techniques to elicit answers with ToM considerations. There have been several papers that have reported success in devising prompting techniques that improve ToM responses (Moghaddam et al., 2023; Wilf et al., 2023; Zhou et al., 2023). Several studies have developed datasets as benchmarks for evaluating ToM in LLMs such as openToM (Xu et al., 2024) and sileod (Sileo et al., 2023).

Our study focuses on examining ToM in LLMs. Specifically, we established a benchmark for evaluation and comparison based on the study by (Van Duijn et al., 2023). We observed that much of the current research in this area relies on prompting methods, much of which yielded promising results. Our goal was to take the benchmark data obtained from (Van Duijn et al., 2023) and evaluate it against models that had been fine-tuned on distinct datasets. This approach allowed us to measure uniformly if ToM had improved on our chosen datasets. We chose datasets that covered crucial aspects of ToM, namely emotion recognition (Rashkin et al., 2018), motivation recognition (Rashkin et al., 2018), false-belief recognition (Le et al., 2019), and epistemic logic (Sileo et al., 2023). We chose to fine-tune on the smaller open-source model Falcon-7b and the larger close-source model GPT-3.5 Turbo.

Our approach seeks to determine whether fine-tuned models, trained on specific datasets can transfer their learned knowledge to a set of questions posed by the (Van Duijn et al., 2023) study. This evaluation method will aid us in understanding if the models possess the capacity to generalize their knowledge of ToM tasks that they were not specifically trained on.

2 Background and Related Work

2.1 Theory of Mind in Humans

“Mentalizing (also called the Theory of Mind) is the ability to explain, predict, and interpret behavior by attributing mental states such as desires, beliefs, intentions, and emotions to oneself and to other people.” (Decety, J et al., 2012). Most humans develop ToM naturally from early childhood to early adolescence, it is believed to be one of the foundational elements of social interactions. There are specific factors that impact the development of this ability resulting in a deficit of ToM. This can lead to difficulties in socializing. The factors could be medical conditions (autism, deafness, and schizophrenia), social factors (family size, and socio-economic factors), or contextual factors (conversational discourse, and aging). (Hofman et al. 2016)

There are four recognized categories of ToM, each one speaks to a different method of social understanding. (Hofman et al. 2016).

First-order false beliefs. First-order false beliefs are the realization that a person can hold false beliefs about events in the world. These false beliefs can be misunderstandings or misconceptions about a situation or fact. For example, John can believe that there is a carton of milk in the fridge because he saw his roommate Sally put it there. However, when John wasn't looking, Sally moved the milk to the freezer. John now holds a first-order false belief about the location of the milk, because he still thinks the milk is in the fridge.

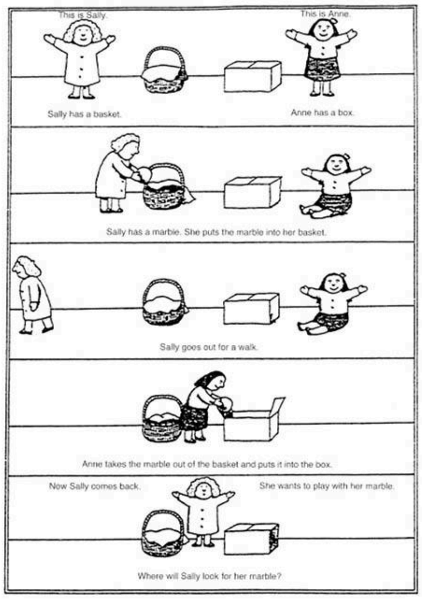
Second-order false beliefs. This is the understanding that a person can hold a false belief about another person's beliefs, even if that belief is wrong. For example, from the previous example, we can see that John believes the milk to be in the fridge because he did not see Sally move it to the freezer. John possesses first-order false belief, however, their other roommate Alice walks in while Sally is putting the milk in the freezer. Now Alice believes that Sally thinks John will look for the milk in the freezer. Alice now holds a second-order false belief, because Sally knows that John will look for the milk in the fridge, but Alice thinks (falsely) that Sally thinks John will look for the milk in the freezer.

Double Bluff. This refers to a situation where someone pretends to do the opposite of what they want to achieve to mislead others who might be trying to predict their actions.

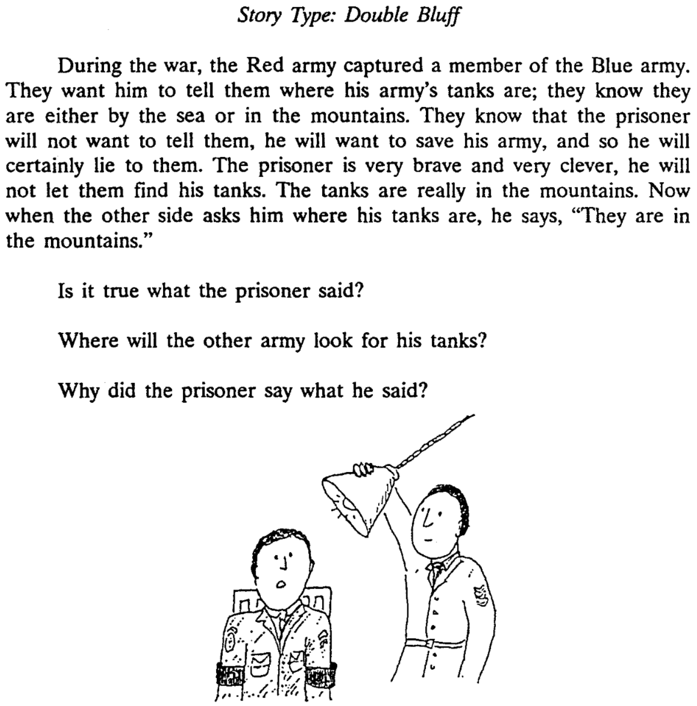
Desire Reasoning. This refers to the ability to understand and predict behaviours based on desires, wants, and preferences.

2.2 The Tests for Theory of Mind

Every person develops their own level of ToM capabilities to a different degree. To measure and evaluate the degree of ToM present in a person, a number of tests were created. The tests are distinguished by the previously mentioned 4 categories of ToM. The most popular and simple test is the Sally-Anne test described in *Figure 1* (*Sally–Anne Test*, n.d.), this test aims at evaluating first-order false beliefs.


This is Sally.
This is Anne.
Sally has a basket.
Anne has a box.
Sally has a marble. She puts the marble into her basket.
Sally goes out for a walk.
Anne takes the marble out of the basket and puts it into the box.
Now Sally comes back.
She wants to play with her marble.
Where will Sally look for her marble?

The correct answer in *Figure 1's* scenario is the basket. A person who possesses ToM capabilities can recognize in this situation that Sally will have a false belief about the location of the marble. She will look for the marble in the basket, even though the marble presides in the box. Generally, most people have no trouble discerning why Sally believes the marble is in the basket. People with medical conditions like autism have been shown in studies to pass first-order false belief tests too. Thus more advanced tests were created to determine categories and precise upon the evaluation of individuals.(Happé, F.G.E., 1994).



Story Type: Double Bluff

During the war, the Red army captured a member of the Blue army. They want him to tell them where his army's tanks are; they know they are either by the sea or in the mountains. They know that the prisoner will not want to tell them, he will want to save his army, and so he will certainly lie to them. The prisoner is very brave and very clever, he will not let them find his tanks. The tanks are really in the mountains. Now when the other side asks him where his tanks are, he says, "They are in the mountains."

Is it true what the prisoner said?

Where will the other army look for his tanks?

Why did the prisoner say what he said?

An example of an advanced test is given in *Figure 2*. It includes a description of a story and a set of questions. To answer these complicated tests correctly a person is required to understand the emotions and motivations of characters as well as the circumstances of a story. It was shown that younger people or people with underdeveloped ToM abilities struggled with these types of questions. (Happé, F.G.E., 1994).

2.3 Theory of Mind in LLM's

The evolution of Large Language Models led to the creation and instant popularity of Chat Generative models capable of long and complex conversations. Such communication can feel like talking to a real individual and so can be considered a social interaction. As established earlier, for humans, ToM is a fundamental part of socialization, but what about LLMs? Recently a few studies have come out researching this topic and generally agree that LLMs have ToM capabilities in some capacity. However, they often struggle and fall behind human capabilities. (Sap et al., 2022), (Xu, H. et al., 2024). The topic was recognized as an important aspect of LLMs, evident from the number of papers that followed. These papers came up with the benchmarks (Xu, H. et al., 2024), (He, Y. et al., 2023) and datasets (Sileo et al., 2023) for evaluation, testing state-of-the-art LLMs on ToM tasks (Xu, H. et al., 2024), (Duijn, et al., 2023) and developing techniques to improve ToM capabilities. (Moghaddam, et al., 2023), (Wilf, et al., 2023).

ToM Evaluation. One recent area of research has been interested in discovering appropriate techniques and datasets to assess the performance of LLMs on tasks that require the application of ToM abilities. The evaluation methodology often consists of asking open-ended questions and grading responses based on human feedback. The tests performed on the LLMs are typically used again for evaluating human capabilities. These tests are focused on assessing the ability to understand and keep track of both the mental and physical states of different agents. Each question includes some context (ie the scenario) a few sentences long describing the characters and their surroundings. For example, in the OpenToM benchmark (Xu, H. et al., 2024), the supporting narrative's goal is to make a clearer story, unraveling the character's personality traits, and hinting at the agent's motives through the description of their actions. Zero-shot setting is a standard practice for evaluations.

Testing state-of-art LLMs. Although papers containing benchmarks for evaluation typically include the results of their tests on some state-of-the-art models, there is more research being done focusing exclusively on deep and detailed analysis of current LLMs. For example, the study by (Duijn, et al., 2023) examines 11 LLMs. These LLMs were tested extensively on a variety of tasks with differing categories and orders of ToM and the results were compared to the performance of children. The results revealed that only a few models with the greatest number of parameters did not struggle with the tests and outperformed the humans.

Improving ToM capabilities. The next logical step after assessing the current state of ToM abilities in LLMs, and concluding that the results show promise but significantly lack behind humans was to attempt to seek an improvement. Most efforts in this area of research are aimed at developing prompting techniques. Some examples of explored approaches include a two-shot chain of thought reasoning, step-by-step thinking instructions (Moghaddam, et al., 2023), and a two-stage prompting framework based on the notion of perspective-taking (Wilf, et al., 2023). When provided with in-context learning prompts LLMs demonstrated an improved performance on ToM tasks showcasing the room for development in base models.

Fine-tuning as ToM enchantment technique. Fine-tuning is a widely recognized method for adapting pre-trained models to some particular task. This process comes in handy when the data for a specific task is limited and when trying to achieve better performance from the base models. It was already shown that given additional context, LLMs can better understand and perform ToM-related tasks. However, there has not been any substantial experimentation on applying a fine-tuning process for this specific ability of LLMs. Our research aims to mitigate this gap in current research of ToM in NLP by conducting fine-tuning over selected datasets developed specifically for this topic. The benefits of this process were evaluated based on benchmarks for the state-of-art models

3 Methodology

In their 2023 paper (Van Duijn et al., 2023) tested the Theory of Mind of 11 LLMs and used responses from a group of children aged 7-10 as a benchmark to compare the performance of the models to. In all but the largest closed-source models, the LLMs underperformed in comparison to the children. Van Duijn et al suggest that some of this may be due to the fact that "…this test relies less on a very specific sort of advanced language ability, but more on a type of behaviourally-situated reasoning…" (Van Duijn et al., 2023). While this may be true studies have shown "…ToM training procedures were effective in improving children's ToM skills…" and "…that in most cases children were presented with situations (in the form of stories, picture books, videos, etc.) …" (Hofman et al., 2016). As these are the kinds of materials that can be used to train LLM's we were inspired to adopt this approach when attempting to improve the ToM of LLMs.

As there have already been numerous studies showing that LLMs were able to improve Theory of Mind of a specific task through specifically designed datasets or prompting techniques (Zhou et al., 2023; Wilf et al., 2023; Moghaddam et al., 2023), we were inspired by earlier breakthroughs in NLP whereby transfer learning was used to improve one area through training in another. For example, improvements in language modeling made advancements in machine translation to see whether training on different ToM datasets would impart a more general ToM into the model which would then transfer to different ToM tests.

3.1 Fine-Tuning

Fine-tuning involves customizing a pre-trained Large Language Model (LLM) for a specific task. While it doesn't entail retraining the model's knowledge base, it does enable it to learn to identify new patterns in data. In this study, our goal is to guide an LLM toward recognizing behavioural patterns that enhance its Theory of Mind (ToM) abilities.

4 Experiments

4.1 Experiment Settings

Implementation. Using the results from the study conducted by Van Duijn et al. (2023), we established a benchmark for comparison. We then tried to replicate their original study as best we could by interrogating the models using the original question set and keeping parameter settings such as temperature, top_p and max_tokens the same but using models which had been fine-tuned using a number of different datasets. We also had only one person scoring each of the models to establish consistency.

Datasets. 4 datasets were selected, each covering a different method of ToM training. All 4 datasets were originally formatted for classification tasks and were modified to be used for inference tasks using some version of the context, prompt, question, answer framework that best fit that dataset. Each model was fine-tuned on individual datasets as well as combinations of the datasets. The datasets created were:

Motivation was adapted from the work by Rashkin et al (2018) and uses stories to understand character motivations. We used 175,000 training examples

Emotion was also adapted from the same dataset at Motivation (Rashkin et al., 2018). These stories were used to extract characters' emotion-driven actions. This training set was also 175,000

ToMi was based on the dataset by Le et al. (2018) and uses questions and answers to perform the Sally Anne test. We used ~6,000 training examples.

Sileod was based on the dataset by Sileod et al. (2023) that uses epistemic logic to test ToM. We used ~10,000 training examples.

Models. Due to restrictions in resources, we were not able to test all the models in the original study, so we opted to test one open and one closed-source model. Falcon 7B was the smaller of the two, it is an open-source base model and GPT 3.5 was the larger closed-source instruct model.

4.2 Experimental Results

Falcon 7B

As seen in table 1 the fine-tuned Falcon 7B models had worse overall performance than the base model with the exceptions being those fine-tuned on the Motivation dataset and the combination of Sileod, ToMi and Emotion.

On the Sally Anne test which was spread over 3 questions table 2 showcases that no model was able to outperform the base model on all parts. However, the combinations of Sileod, ToMi, and Emotion; Sileon, ToMi, and Motivation; and the All combination outperformed the base model on question 2 and Sileod, ToMi, Emotion, and All combination models also outperformed the base model on questions 3. Question 2 deals with either first-order or second-order beliefs. Question 3 requires reasoning on the statement in question 2. The base model was unable to answer any of the 3rd part of this test correctly.

Table 3 shows the outcome from the Strange Stories section with the base model outperforming all other models on the initial question which dealt with the task of identifying whether somebody had told a lie, a joke, or was using sarcasm. In all cases, the only scenario the models were able to identify was lying. In the 2nd part of the test, we asked the model to give the reason for the lie, joke, or sarcasm, all but the ToMi model outperformed the base model

The final test, Imposing Memory, was poorly performed across all models. This question included a long story and required the model to decide whether the final statement was correct or incorrect. In almost all cases the models defaulted to outputting [correct] as the answer.

GPT 3.5

Table 1 presents our overall performance findings, if we observe the findings from all the GPT-3.5 Turbo models, we can see that all models perform worse than the original base instruct model. However, we can observe that most of the GPT models perform better or just as well as their Falcon counterparts.

In Table 5, we can see the results from the Sally Anne Test, one of the most popular methods for testing ToM. Interestingly not a single model was able to outperform the base GPT model. The models showcased difficulty in both identifying the right answer and providing reasoning for the answer.

The second staging of questions came from the Strange Stories section. Table 6 shows the results of this line of questioning, we can see that the models again did not perform better than the base GPT model. This table shows the average of the model's answers to two questions. The model consistently, answered the first question wrong (identify lying, joking, or sarcasm), but the model almost always was able to reason why the person was lying, joking, or displaying sarcasm.

The final test was the Imposing Memory test. As described in the Falcon results, the model was given a long scenario and a question, the model had to answer with a yes or no answer. Table 7 shows the results from this final stage of testing. Both the ToMi and sileod models performed better than the base GPT model, but the motivation underperformed compared to the base model.

5 Discussion

Falcon 7B

The outcomes of these tests were interesting. With regards to Falcon 7B, there was no change at all with many models creating almost the exact output given by the base model with no deviation in word choice or sentence structure. However, one of the most striking differences was produced by the ToMi model which produced almost entirely hallucinatory outputs which drew heavily on vocabulary from the ToMi dataset which was also the smallest dataset. We have to assume that the vocabulary in ToMi deviated greater from that in the final questions than the vocabulary in the other datasets which suggests that when fine-tuning a model new vocabulary is somehow weighted much greater than existing vocabulary weights. Something similar was noted in the Emotion model, vocabulary from its dataset was transposed into answers resulting in hallucinations [figure 1]. In contrast, the Motivation model performed well and yet the Emotion and Motivation datasets were abstracted from the same original dataset. It should be noted that the wording in the Emotion dataset is a little more awkward and perhaps less rich than in the Motivation dataset. The abstracted version of the Emotion dataset asks "Did [character] have an emotional reaction in Line [number]?" Perhaps a better training set would have included what the emotion was, this data was available in the original dataset. This could have improved the answers from being able to identify a general emotional reaction to specifying the emotion.

GPT-3.5 Turbo

The overall performance of our GPT models performed on average better than the Falcon models. In general, the ToMi and Sileod models performed better than the motivation model. ToMi was the better model of the two, in comparison to the Falcon models, there was no hallucination observed in its outputs. Suggesting that something in the Falcon model caused the hallucinations to happen. Interestingly, however, the motivation dataset had hallucinations within its outputs, these outputs whilst they could make sense in the right context, did not, in this one. The Sally Anne Test, the ToMi model seemed to just copy what the golden answer was from the first question across the first two questions posed to it. These questions were based on first and second-order belief. The sileod dataset model on the other hand always got the first-order belief question right, but on the other too, could not identify the second-order belief or the reasoning behind it, which would logically make sense. An interesting pattern across all models was encountered regarding the Strange Stories series of questions. Strange Stories is split into 7 specific sections, all models struggled to identify if the questions posed to them had a misunderstanding in them. This could be an area to further research and improve on. Another observation that we encountered was that the models on average could better explain the reasoning behind a question and get it correct, but more often than not they could not identify whether the question was true or false, or provide the correct identification word.

In summary, it must be noted that the outputs from these tests were as a result of training models on data different from that in the test questions to ascertain whether a general ToM could be imparted on models through transfer learning and in some cases there was an improvement specifically in the areas of reasoning and explanation.

For future research it may be interesting to look more closely at the structure of the training data, when fine-tuning the information provided is clearly given greater weight in the model and therefore the smallest deviations in the training data can have greater effect on the output.

6 Conclusion

Our experiment indicates that we were only able to produce two models that outperform our base models of Falcon 7b and GPT-3.5 Turbo. However, we had promising results that could lead to further research within the project. There were some specific instances within our Falcon 7b fine-tuning process that yielded results that outperformed the base model, these were specifically when we combined the datasets Sileod, ToMi, and Emotion, and when we fine-tuned on the motivation dataset. Both datasets are rich in text suggesting that the models need bigger datasets to perform well against the ToM tests. Furthermore, with the smallest dataset, we found hallucinations, which would indicate that bigger datasets should be used when fine-tuning. Our GPT models gave insight into the fact that they struggle with identifying specific tasks but have little trouble reasoning what the task was lying, joking, etc. Suggesting that perhaps during training more text with specific labels could improve this aspect of the models. For future studies, working with more varied text-rich datasets, we believe, could improve the results of the models. Additionally, working with more combinations of datasets with the GPT-3.5 Turbo model could yield interesting results as, due to resource restrictions, we were only able to fine-tune using a limited selection of datasets.

Limitations

Although care was taken to replicate the initial study there are factors that will affect the outcomes differently such as the quantizing of models to perform fine-tuning as well as how fine-tuning data is used within the larger model during use.

As with the original study we were only able to have one person scoring each model it may have been better to have several people testing the models and taking averages of the scores. Similarly, we can not say for certain whether we have scored in exactly the same manner as the original study.

References

^ Hofmann, S. G., Doan, S. N., Sprung, M., Wilson, A., Ebesutani, C., Andrews, L., Curtiss, J., & Harris, P. L. (2016, February 20). *Training children's theory-of-mind: A meta-analysis of controlled studies*. NCBI. Retrieved April 15, 2024, from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4792776/

^ Decety, J., & Svetlova, M. (2012). Putting together phylogenetic and ontogenetic perspectives on empathy. *Developmental Cognitive Neuroscience*, *2*(1), 1-24. https://doi.org/10.1016/j.dcn.2011.05.003

^ *Sally–Anne test*. (n.d.). Wikipedia. Retrieved April 23, 2024, from https://en.wikipedia.org/wiki/Sally%E2%80%93Anne_test

^ Happé, F.G.E. An advanced test of theory of mind: Understanding of story characters' thoughts and feelings by able autistic, mentally handicapped, and normal children and adults. *J Autism Dev Disord* **24**, 129–154 (1994). https://doi.org/10.1007/BF02172093

^ Maarten Sap, Ronan Le Bras, Daniel Fried, and Yejin Choi. 2022. Neural Theory-of-Mind? On the Limits of Social Intelligence in Large LMs. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 3762–3780, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Xu, H., Zhao, R., Zhu, L., Du, J., & He, Y. (2024). OpenToM: A Comprehensive Benchmark for Evaluating Theory-of-Mind Reasoning Capabilities of Large Language Models. *ArXiv*. /abs/2402.06044

^ He, Y., Wu, Y., Jia, Y., Mihalcea, R., Chen, Y., & Deng, N. (2023). HI-TOM: A Benchmark for Evaluating Higher-Order Theory of Mind Reasoning in Large Language Models. *ArXiv*. /abs/2310.16755

^ Sileo, D. and Lernould, A., 2023. Mindgames: Targeting theory of mind in large language models with dynamic epistemic modal logic. *arXiv preprint arXiv:2305.03353*.

^ Max van Duijn, Bram van Dijk, Tom Kouwenhoven, Werner de Valk, Marco Spruit, and Peter van der Putten. 2023. Theory of Mind in Large Language Models: Examining Performance of 11 State-of-the-Art models vs. Children Aged 7-10 on Advanced Tests. In *Proceedings of the 27th Conference on Computational Natural Language Learning (CoNLL)*, pages 389–402, Singapore. Association for Computational Linguistics.

^ Moghaddam, S.R. and Honey, C.J., 2023. Boosting Theory-of-Mind Performance in Large Language Models via Prompting. *arXiv preprint arXiv:2304.11490*.

Wilf, A., Lee, S.S., Liang, P.P. and Morency, L.P., 2023. Think Twice: Perspective-Taking Improves Large Language Models' Theory-of-Mind Capabilities. *arXiv preprint arXiv:2311.10227*.

^ Le, M., Boureau, Y.-L., & Nickel, M. (2019). Revisiting the Evaluation of Theory of Mind through Question Answering. *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 5872-5877. https://aclanthology.org/D19-1598/

^ Rashkin, H., Bosselut, A., Sap, M., Knight, K., & Choi, Y. (2018). Modeling Naive Psychology of Characters in Simple Commonsense Stories. https://uwnlp.github.io/storycommonsense/data/rashkin2018modeling.pdf

Zhou, P., Madaan, A., Potharaju, S. P., Gupta, A., McKee, K. R., Holtzman, A., ... & Faruqui, M. (2023). How FaR Are Large Language Models From Agents with Theory-of-Mind?. *arXiv preprint arXiv:2310.03051*.

Rashkin, H., Bosselut, A., Sap, M., Knight, K., & Choi, Y. (2018). Modeling naive psychology of characters in simple commonsense stories. *arXiv preprint arXiv:1805.06533*.

Appendix

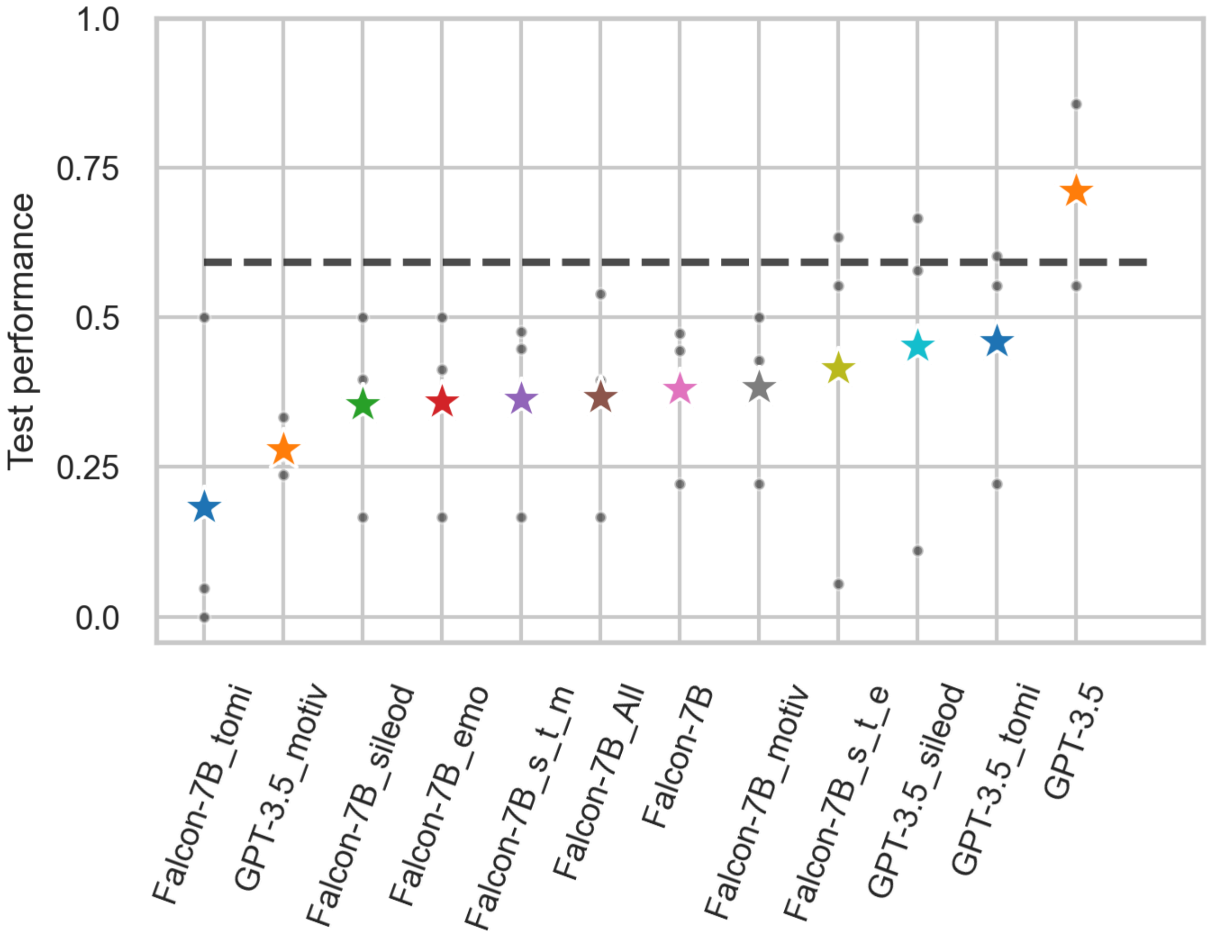


Table 1

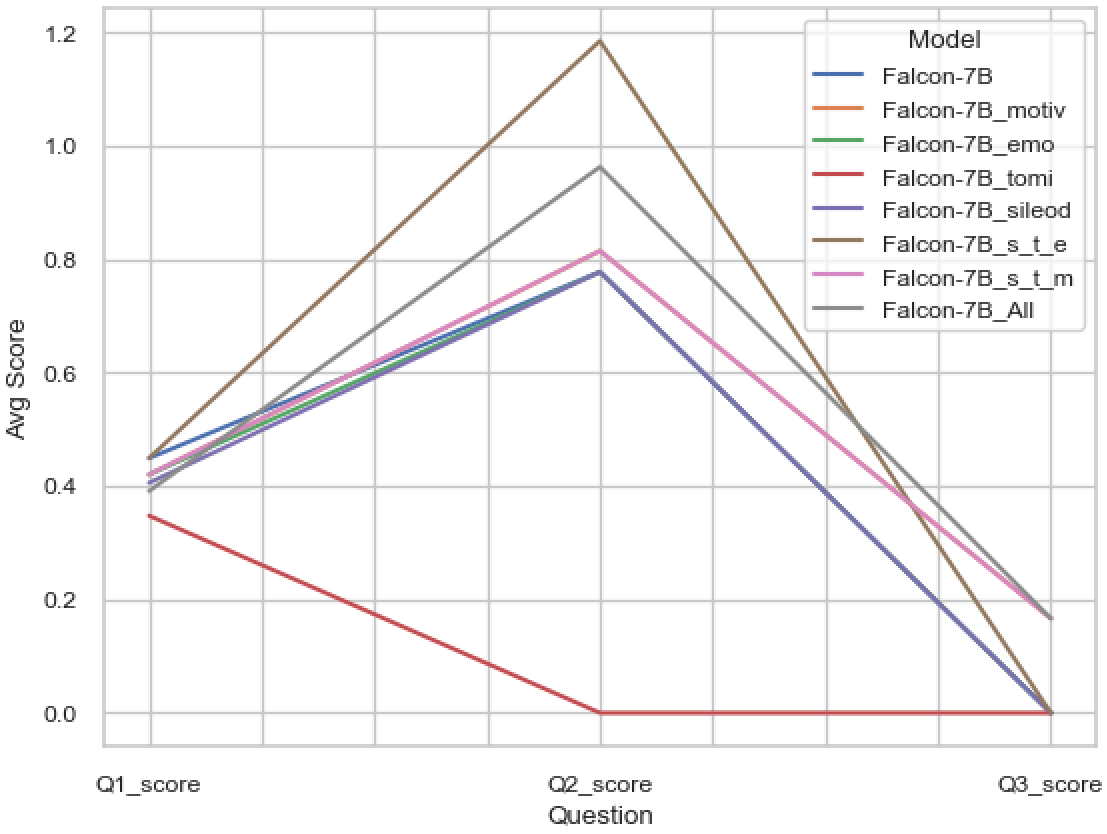

Model
Falcon-7B
Falcon-7B_motiv
Falcon-7B_emo
Falcon-7B_tomi
Falcon-7B_sileod
Falcon-7B_s_t_e
Falcon-7B_s_t_m
Falcon-7B_All
Avg Score
1.2
1.0
0.8
0.6
0.4
0.2
0.0
Q1_score
Q2_score
Q3_score
Question

Table 2

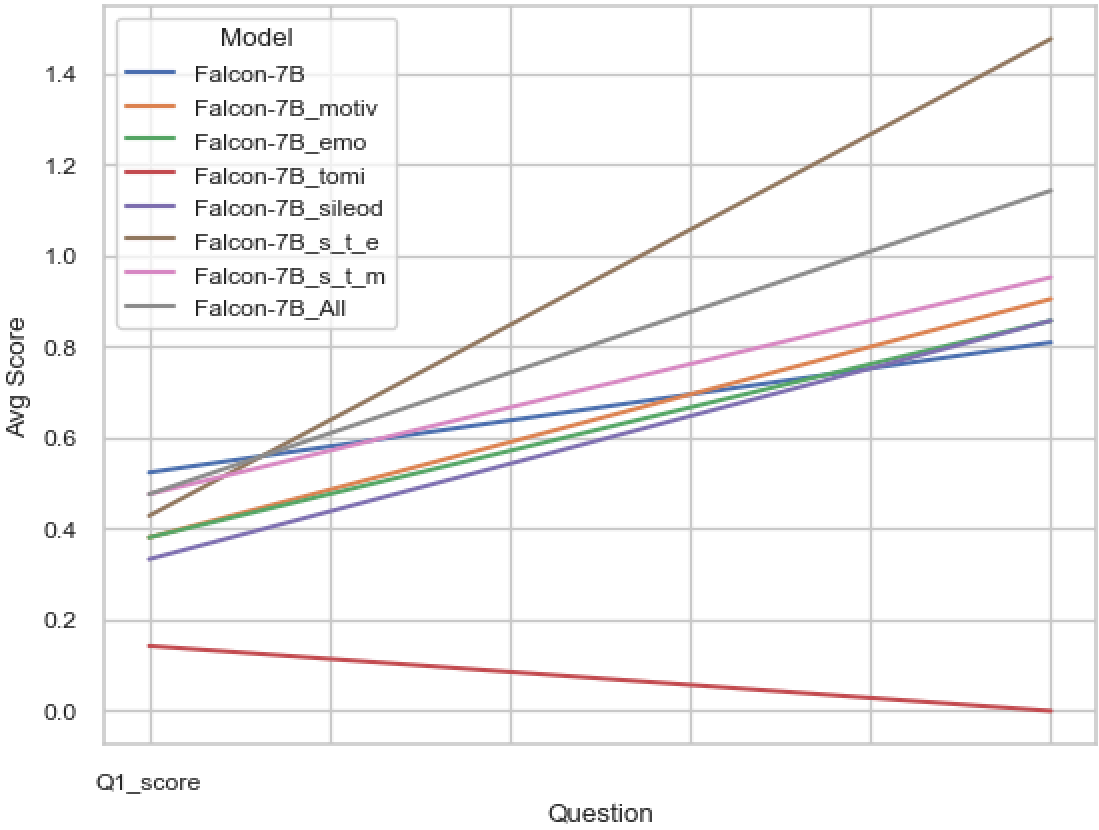

Model
Falcon-7B
Falcon-7B_motiv
Falcon-7B_emo
Falcon-7B_tomi
Falcon-7B_sileod
Falcon-7B_s_t_e
Falcon-7B_s_t_m
Falcon-7B_All
Avg Score
1.4
1.2
1.0
0.8
0.6
0.4
0.2
0.0
Q1_score
Question

Table 3

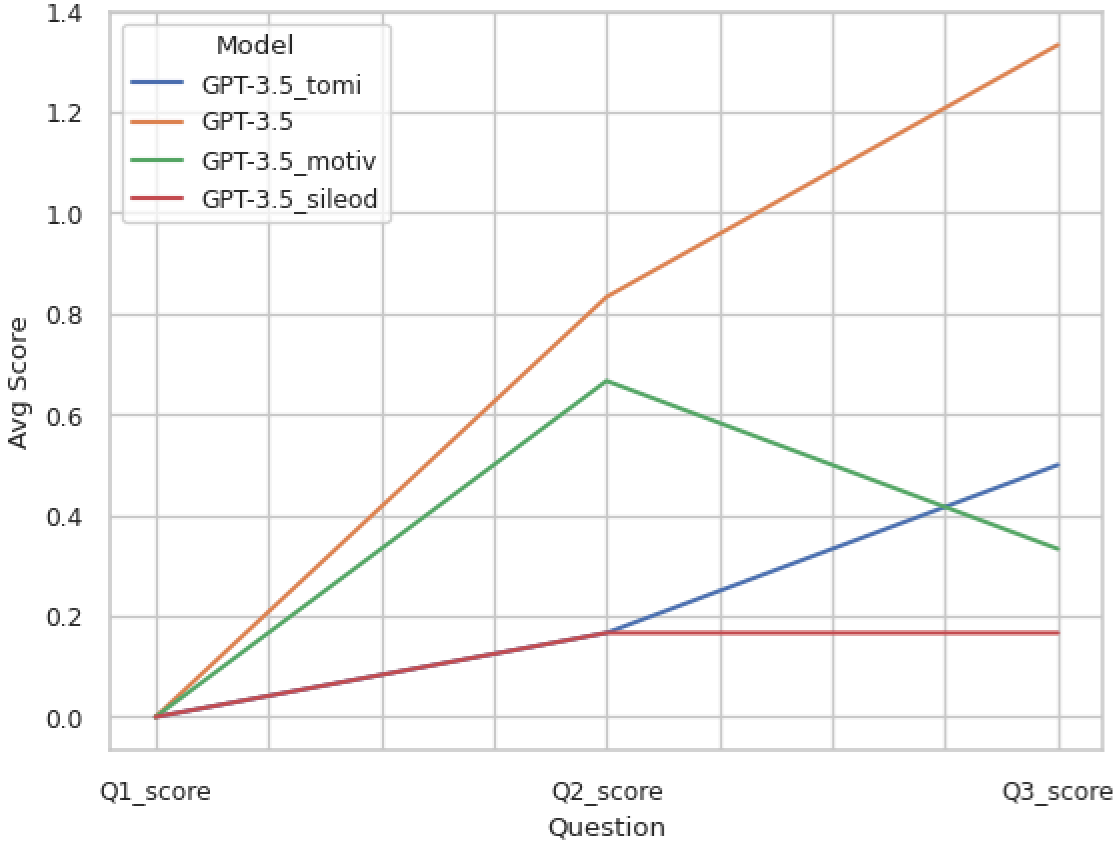

1.4
1.2
1.0
0.8
0.6
0.4
0.2
0.0
Avg Score
Model
GPT-3.5_tomi
GPT-3.5
GPT-3.5_motiv
GPT-3.5_sileod
Q1_score
Q2_score
Q3_score
Question

Table 5 SA

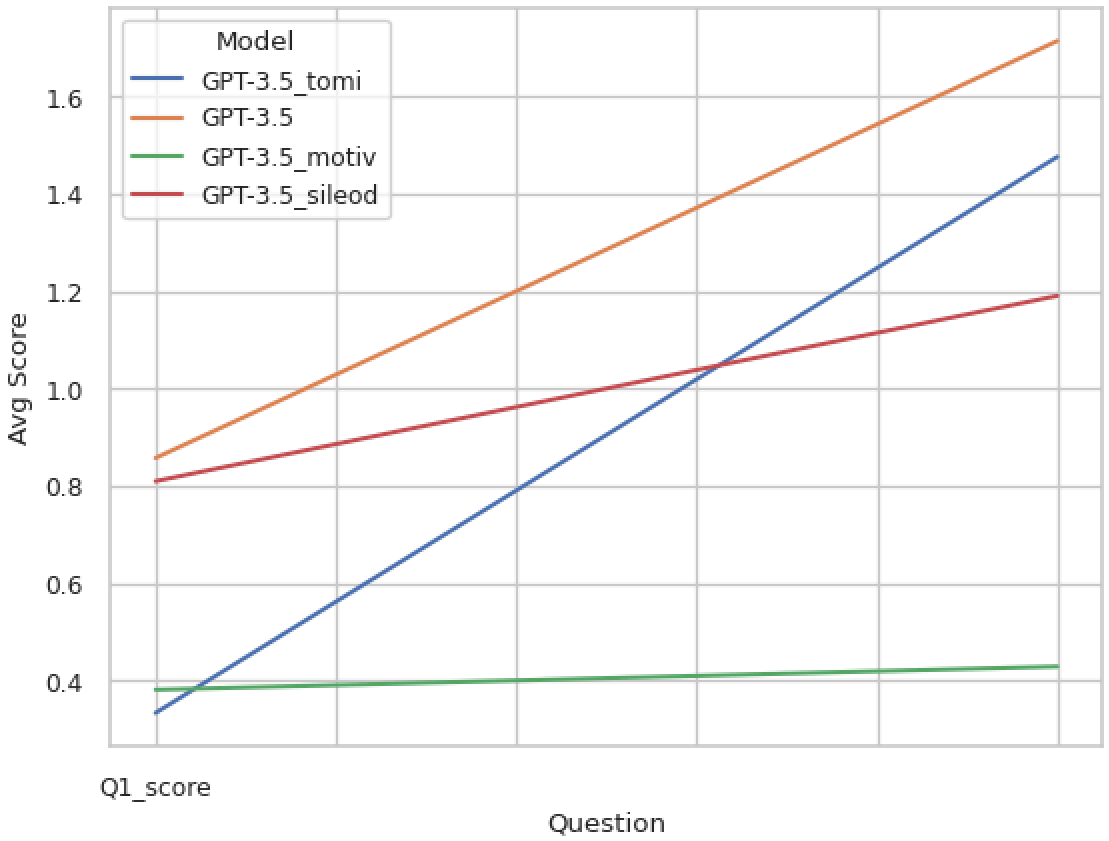

Model
GPT-3.5_tomi
GPT-3.5
GPT-3.5_motiv
GPT-3.5_sileod
1.6
1.4
1.2
1.0
0.8
0.6
0.4
Avg Score
Q1_score
Question

Table 6 ss

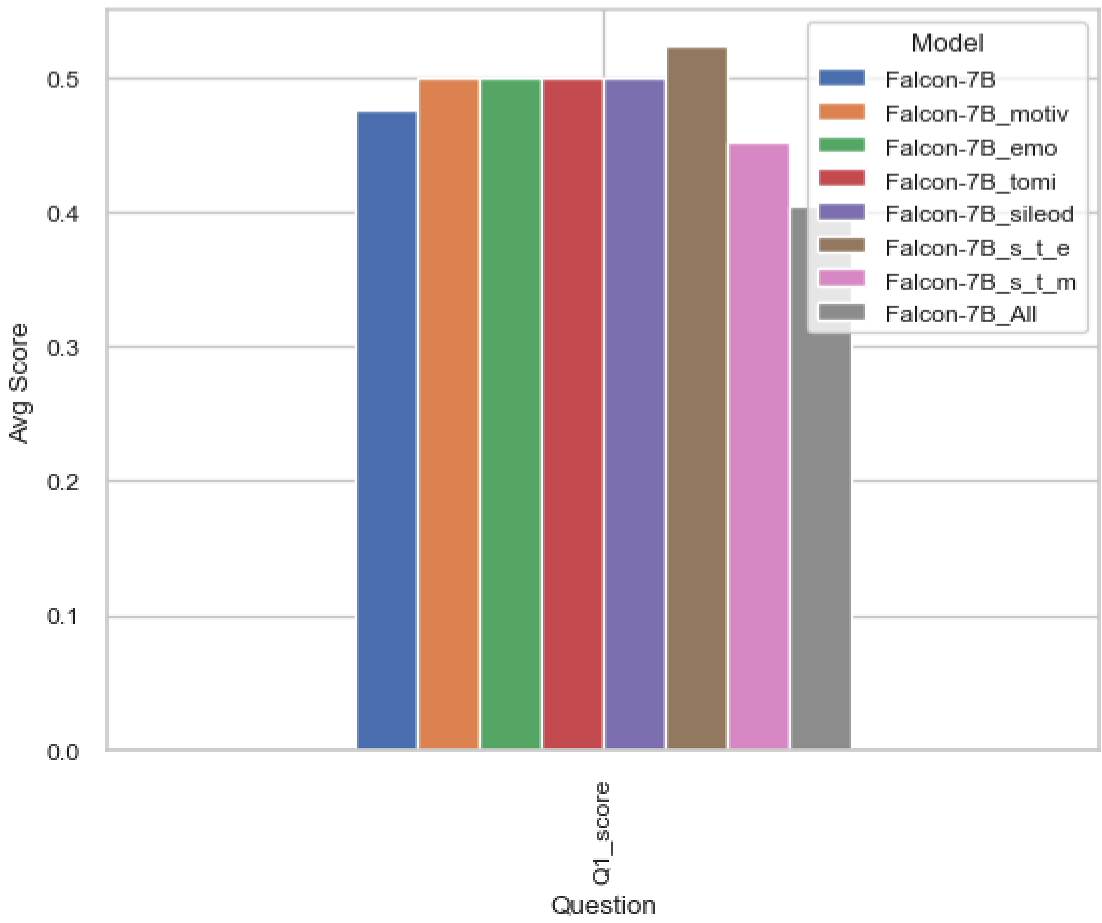

Model
Falcon-7B
Falcon-7B_motiv
Falcon-7B_emo
Falcon-7B_tomi
Falcon-7B_sileod
Falcon-7B_s_t_e
Falcon-7B_s_t_m
Falcon-7B_All
Avg Score
0.0
0.1
0.2
0.3
0.4
0.5
Q1_score
Question

Table 4

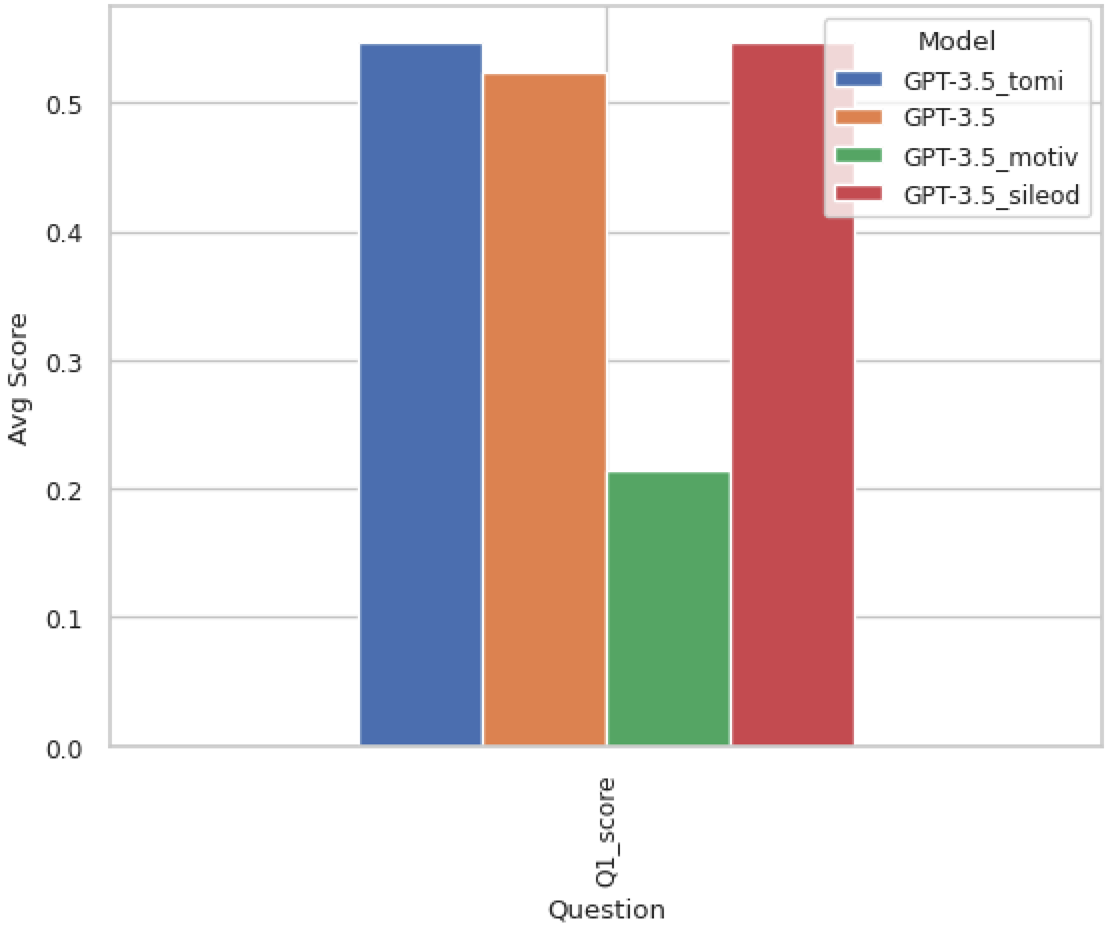


Table 7

Dataset Training Example

Context: A cook was carrying an armful of oranges in the kitchen. He dropped one on the floor by accident. As a joke he pretended that it was a soccer ball. He kicked the orange across the kitchen.

Question: Did Cook have an emotional reaction in line 4

Answer: yes

Output from Model

Context: James goes to Claire's house for the first time. He goes there for a drink and wants to see Claire's dog whom she always talks about. James likes dogs very much. When James gets to Claire's house, Claire runs to the door to open it, and her dog jumps up to greet James. Claire's dog is huge, and almost as tall as James! When James sees Claire's big dog he says, "Claire, you don't have a dog at all. You have an elephant!

Prompt: What James says is

Answer: What James says is an emotional reaction in line 3. James has an emotional reaction in line 4. James has