

Understanding Model Behavior in Tiered Reasoning for Intuitive Physics

Tsedeniya Amare
University of Michigan
tamare@umich.edu

Joan Nwatu
University of Michigan
jnwatu@umich.edu

Abstract

Large pre-trained language models consistently top the leaderboard for benchmarks across a variety of NLP tasks. However, due to the high complexity of these models, it is challenging for researchers, developers, and AI end-users to understand how these models make prediction decisions. Understanding model behavior is crucial for identifying the limitations of AI systems, diagnosing flaws in model architectures and datasets, and building trust with the global community who are affected by AI algorithms in their daily lives. In this work, we leverage the Tiered Reasoning for Intuitive Physics dataset and framework to investigate model behavior on the task of physical commonsense reasoning. Using an explainability tool, LIME, and a prompting approach, we conduct analysis to uncover model behavior in story classification. Our findings identify several factors that affect model decisions and raise some topics for consideration when designing a training pipeline and benchmarks for the task of commonsense.

1 Introduction

It is critical for large language models(LM) to have true reasoning that is comparable to human reasoning and real-world commonsense since humans interact with this LM in various applications area of natural language processing(NLP). However, in the Tiered Reasoning for Intuitive Physics(TRIP) work [Storks et al. \(2021a\)](#), they have shown that large language models (BERT, RoBerta, DeBERTa) perform poorly in language understanding when extracting commonsense inferences. Based on this work we come up with a research question “why do those large language models perform poorly when they are trained to have commonsense reasoning for the end task prediction for the TRIP dataset?” In the paper, they performed different analyses of modalities like accuracy, consistency, and verifiability of different model predictions which shows

how these modalities change across those models when they are trained with commonsense reasoning.

In this project which is an extension of the previous project [Storks et al. \(2021a\)](#), we investigated the behavior of a large language model on the TRIP dataset using an explainability tool which is Interpretable Model-Agnostic Explanations (LIME)[Ribeiro et al. \(2016\)](#). We used the Tiered Reasoning for Intuitive Physics (TRIP) dataset, which includes a high-level end task for story plausibility classification as well as dense annotations for each story that capture multiple tiers of reasoning beyond the end task. We trained BERT model using the story classification part of the dataset to predict plausible and implausible stories. The accuracy of this Bert model is 71 percent which we mainly didn’t focus on improving the accuracy instead we evaluated how our model is learning the data and what feature it is looking at. In particular, we used LIME for one data point to understand what features attributed to the BERT model predictions. We generated the ten most important features using LIME and LIME listed them in form of the feature importance. From our analysis, we learned that the model is trying to learn positive and negative words in the story to predict the plausibility instead of the time series relationship of the physical states of the stories stated in the two conflicting sentences. The model is also emphasizing less important words like an article(e.g the)in the story classification.

In addition to the LIME analysis, we made use of a prompt-dialog guided method to investigate the behavior of GPT-3, a pre-trained language model optimized for few-shot learning, on the TRIP dataset. This method enabled us to probe and identify possible reasons why language models struggle with verifiable common sense reasoning. Findings from this approach can provide feedback for the improvement of common sense language

model architecture and benchmarks.

The contribution of this work is first we implemented LIME in non-probabilistic models like transformers which need additional improvement to implement the LIME on BERT model. For the second contribution, we introduced a new approach to further scrutinize language model behavior on commonsense tasks. Thirdly, from the analysis of model behavior on commonsense tasks, we identified several issues for consideration when designing model pipelines and benchmarks for commonsense reasoning.

2 Related work

There has been a long-standing argument that language models cannot learn commonsense [Davis and Marcus \(2015\)](#). While some research [Trinh and Le \(2019\)](#) assert that language models are capable of extracting some commonsense knowledge by predicting how probable a piece of text is in a given context, there are numerous reports of language models and their inability to produce acceptable results across various commonsense benchmarks [Kejriwal and Shen \(2020\)](#); [Bisk et al. \(2019\)](#); [Storks et al. \(2021a\)](#).

Another approach that involves the creation of commonsense rules to populate a knowledge base has been found inapplicable to natural language processing tasks due to the inconsistency and complexity of information retrieval [Mahesh et al. \(1996\)](#). A more efficient approach involves using deep pre-trained language models finetuned on seed knowledge bases such as [Sap et al. \(2018\)](#); [Speer et al. \(2016\)](#) to generate new commonsense representations [Bosselut et al. \(2019\)](#).

However, if language models developed through unsupervised learning from free-form text are able to show excellent results in performing specific end tasks like question answering, we can expect the model to be able to correctly identify some evidence within the text that supports its prediction. The results in [Storks et al. \(2021a\)](#) indicate that this is not the case for transformer-based language models. This generates suspicion that language models do not truly understand language to the extent that is sufficient for them to successfully perform tasks like common sense reasoning, and any excellent performance from the models on such tasks can be attributed to the models' ability to exploit statistical cues within the data. To probe this suspicion we seek to understand model behavior.

2.1 Explainability tool: LIME

We used an explainability tool to investigate the results of the transformer-based models. In recent years, researchers have focused on understanding machine learning's black-box behavior and justifying their decisions to users using explainability Tools. Furthermore, explainability tools can be used in addition to metrics evaluation such as accuracy [Leiter et al. \(2022\)](#). These explainability tools can validate different models' context capture capabilities [Chhipa et al. \(2020\)](#) And capture meaningful patterns diving model behavior, allowing developers to 'plug and play' to better model patterns [Szczepański et al. \(2021\)](#).

2.2 Prompting for Generation of Model Explanations

Another approach used by researchers to investigate model behavior involves allowing the language models to generate human-understandable explanations for their predictions using natural language like English.

Following the self-talk paper ([Shwartz et al., 2020](#)) which implemented a generational approach to common sense as inspired by inquiry-based discovery learning (Bruner, 1961), common sense research moved towards the direction of extracting relational information from language models. ([Petroni et al., 2019](#)) conducted a study to compare information extraction from knowledge bases to the performance of BERT (without finetuning) on cloze queries for factual and commonsense information. A popular and effective method for extracting information is the use of prompts. ([Brown et al., 2020](#)) demonstrated that by providing a piece of text, usually referred to as a prompt, before the original task in as little as a few examples, models like GPT-3 can achieve performances that are competitive with fine-tuning approaches in downstream tasks. A major advantage of prompting is that a few examples can be used to condition a pre-trained language model to perform well on a downstream task, contrary to the huge number of training examples required for finetuning to yield similar results. ([Ahn et al., 2022](#)) used prompts to investigate the ability of large language models to plan and break down grounded instructions using levels of feedback injected into the prompt. Such tasks demand the pre-trained model to draw from the relational world knowledge stored within the model and apply 'reasoning' to generate a coherent response to

the task to agent planning. A method introduced by (Wei et al., 2022), called Chain of thought prompting enables language models to learn from examples with a series of intermediate steps on how to perform complex reasoning tasks. This method has improved the efficacy of prompting for arithmetic, commonsense, and symbolic reasoning tasks.

(Camburu et al., 2020) showed that language models are capable of generating human-understandable explanations of their predictions in natural languages such as English. While some of these explanations tend to be inconsistent, some of them expose flaws in the decision-making process of the models. To address the same issue of generating explanations, (Paranjape et al., 2021) uses the idea of counterfactuals to prompt language models into providing contrastive explanations for their decisions.

Similarly, our work uses prompting to generate understandable explanations for model decisions on the TRIP dataset, however, the prompting style incorporates a chain-of-thought approach.

3 Dataset

For this project, we will make use of the Tiered Reasoning for Intuitive Physics (TRIP) dataset which is available at Storks et al. (2021b). The dataset contains multi-level annotations that capture commonsense reasoning. Each training example contains two stories with an equal number of sentences. Both stories are similar except for one sentence where they differ, with one story having a sentence that disrupts the plausibility of the story. Systems can be trained on this dataset to identify the following;

- The plausible story
- The pair of sentences that cause the conflict of implausibility in the unlikely story
- The inherent physical states present in the sentences identified as responsible for the conflict

As humans are able to display commonsense by supporting their argument for choosing one story as unlikely over the other with a description of the exact sentences and states that cause conflict in the story flow, we can check if machine predictions come from sound reasoning by testing if they can support their predictions with the right evidence using the TRIP dataset.

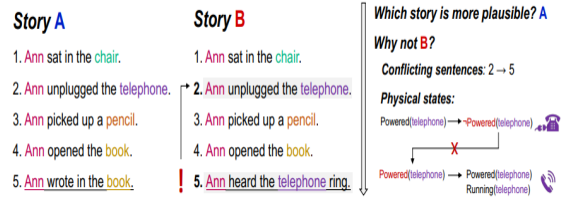


Figure 1: Story pair from TRIP, along with the tiers of annotation available to represent the reasoning process.

(Storks et al., 2021a)

4 Approaches

4.1 Using Explainability tools

We used LIME to understand the behavior of the LM model on TRIP. The motivation for this direction is that large language models are black boxes and we do not know what is going on inside them. The LIME tool can help us to unpack the nature of the models and understand the model's learning patterns. We examined the reasons behind the prediction of these different models stated in the paper. The other motivation is also to see what kind of reasoning these models exhibit.

4.1.1 Interpretable Model-Agnostic Explanations (LIME)

LIME is an explainable AI tool that helps us with Local Interpretability of the model. Which answers the question, "Why is the model behaving in a specific way in the locality of a data point x ?". LIME is a model that takes the probabilistic output of the model prediction. After we pass the probability prediction to the LIME model, LIME generates five thousand nearby data points to that specific data point and outputs the feature attribution for the specific prediction.

4.1.2 TRIP Data Preparation

1. **Story Classification :** The TRIP dataset has three parts: one is the story classification with five sentences. The second part is physical state classification and the third is conflicting sentence detection. We used the story classification part of the dataset which we preprocessed into a sequence of five sentences and used these sentences as sequential input for BERT pretrained model. The input for the BERT model is the stories with five sentences and the output for BERT model is implausible or plausible.
2. **LIME data:** For LIME we change the binary output of the BERT model into the probability

of the story’s plausibility and implausibility. The output of LIME is a list of explanations, reflecting the contribution of each feature to the prediction of a data sample.

4.1.3 Model Training

We fine-tuned BERT (Devlin et al., 2018) large language model using the story classification data which is trained on training data and evaluated using the testing dataset from TRIP. For fine-tuning, we used the data on the GitHub page for the TRIP paper Storks et al. (2021a). This data was mainly preprocessed for the multi-Tiered fine-tuning setup, we further processed the data in paragraph format to use it for the BERT model. For fine-tuning, we used different hyperparameters for example optimizer, batch size, epochs, and different sizes of BERT pre-trained models. We got an accuracy of 71 % 1 for testing purposes for LIME.

Large language Model	Accuracy
BERT	71%

Table 1: Accuracy of BERT model

4.1.4 Testing With LIME

To implement LIME, we convert the binary output of the BERT model into the probability of the plausibility or implausibility of the story. This means instead of the story classification we passed the probability of a story to be plausible or implausible.

LIME generates a list of explanations that reflect the contribution of each feature to the prediction of a data sample this is shown in figure 1. This allows for local interpretability while also determining which feature changes will have the greatest impact on the prediction.

4.2 Prompt guided dialog

We applied a dialog approach for providing input data as a prompt to GPT-3 which generates a prediction about the input data. Questions about the model decision are added to the generated output and fed into GPT-3 to explore the model’s reasoning for the prediction and generate an explanation.

4.2.1 GPT-3

GPT-3 is an autoregressive language model that is capable of high performance on many NLP datasets, such as question-answering, cloze tasks, and tasks that require reasoning or domain adaptation (Brown et al., 2020). In contrast to the fine-

tuning approach that requires thousands of examples and gradient updates of a billion parameters, GPT-3 learns from few-shot examples of the task provided as textual interactions to the model.

4.2.2 TRIP Data Preparation

1. **Chain of thought Prompt Adaptation:** (Wei et al., 2022) shows that enriching examples with a series of intermediate steps on how to perform tasks improves the language model’s few-shot performance on complex reasoning tasks. Therefore, we adapted a pair of example stories from TRIP; a plausible story and an implausible story, to include assumptions about each story. These assumptions serve as intermediate steps of reasoning that guide the model toward making the right prediction.
2. **Story Adaptation:** For each story example, we ordered the sentences in each story to indicate the chronological order of the events.

4.2.3 Example Responses

We present example responses in the form of a human-bot conversation below. The first six interactions make up the initial prompt, including few-shot examples of story classification. After the sixth interaction, the model is prompted to give a response to the last interaction from the human following the guided dialog pattern used in the prompt. The goal of the model in this case is to classify the last story presented to the model.

Robot: Hi there, I’m a robot designed to infer implicit information from short stories.

Robot: You can tell me a story and I will tell you the assumptions that can be derived from the story through commonsense and tell you if the story is plausible.

Human: 1. Ann sat in the chair. 2. Ann unplugged the telephone. 3. Ann picked up a pencil. 4. Ann opened the book. 5. Ann wrote in the book

Robot: 1. Ann is a person. A person can sit on a chair. A person can unplug a telephone. A person can pick up a pencil. A person can open a book. A person can write in a book. 2. A chair can be sat on by a person. If a person can sit on the chair, the chair is not broken 3. A telephone can be unplugged. An unplugged telephone cannot receive incoming phone calls. An unplugged telephone can be used to call another person. 4. A pencil can be used by a person to write 5. A book can be opened by a person. A pencil can be used to write in a

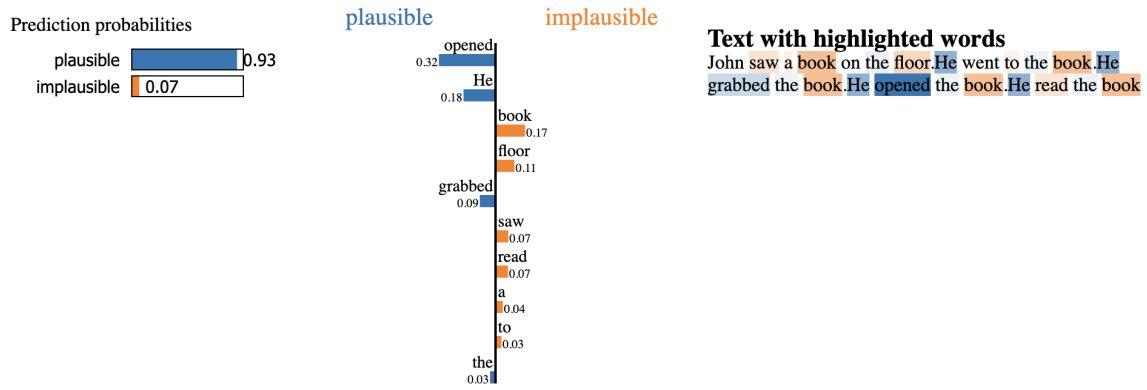


Figure 1: Sample LIME output

Robot: Hi there, I'm a robot designed to infer implicit information from short stories.
 Robot: You can tell me a story and I will tell you the assumptions that can be derived from the story through commonsense and tell you if the story is plausible.
 Human: 1. Ann sat in the chair. 2. Ann unplugged the telephone. 3. Ann picked up a pencil. 4. Ann opened the book. 5. Ann wrote in the book
 Robot: 1. Ann is a person. A person can seat on a chair. A person can unplug a telephone. A person can pick up a pencil. A person can open a book. A person can write in a book. 2. A chair can be sat on by a person. If a person can seat on the chair, the chair is not broken 3. A telephone can be unplugged. An unplugged telephone cannot receive incoming phone calls. An unplugged telephone can be used to call another person. 4. A pencil can be used by a person to write 5. A book can be opened by a person. A pencil can be used to write in a book. 6. Everything makes sense so this story is plausible.

Figure 2: Example of Chain of Thought adaptation with the assumptions highlighted in yellow

book. 6. Everything makes sense so this story is plausible.

Human: 1. Ann sat in the chair. 2. Ann unplugged the telephone. 3. Ann picked up a pencil. 4. Ann opened the book. 5. Ann heard the telephone ring.

Robot: 1. Ann is a person. A person can sit on a chair. A person can unplug a telephone. A person can pick up a pencil. A person can open a book. A person can write in a book. 2. A chair can be sat on by a person. If a person can sit on the chair, the chair is not broken 3. A telephone can be unplugged. An unplugged telephone cannot receive incoming phone calls. An unplugged telephone can be used to call another person. 4. A book can be opened by a person. 5. Ann heard the telephone ring after she unplugged it. Since an unplugged telephone cannot receive incoming calls, this story does not make sense. The story is implausible.

Human: 1. Tom bought a new dustbin for the kitchen. 2. Tom threw a broken plate in the dustbin. 3. Tom got some soup from the fridge. 4. Tom put

the soup in the microwave. 5. Tom ate the cold soup.

Robot: 1. Tom is a person. A person can buy a dustbin. A person can throw a broken plate in a dustbin. A person can get soup from a fridge. A person can put soup in the microwave. A person can eat soup. 2. A dustbin can be bought by a person. A broken plate can be thrown in a dustbin. Soup can be taken from a fridge. Soup can be put in a microwave. Soup can be eaten by a person. 3. Everything makes sense so this story is plausible.

Human: The story is actually implausible. Can you find out your mistake?

Robot: Yes. The story was implausible because it was impossible for Tom to eat cold soup after he put it in the microwave.

5 Results and Evaluation

5.1 Results from LIME

We used LIME to test several data points from the TRIP test data. Because LIME only tests one data point at a time, we were unable to test all of them.



Figure 3: Notable results from LIME mode. Trends are discussed in Section 5.1

However, the evaluation we performed yielded consistent results from the data points we tested. The first is that the fine-tuned BERT model is learning a pattern of positive and negative words so that the model predicts plausible if the story contains positive words and implausible if the story contains negative words. This is illustrated in figure3, where LIME ranked the word "open" as the most important word in the story for model prediction. The model correctly predicts the plausible story, but because the model emphasizes the word "open," it incorrectly predicted the second story as plausible when it was originally labeled as implausible. To support our observation, we change the word "open" to "close," which are opposites, and the story becomes implausible, which does not make sense.

The second evaluation was to see how the positive word affected the prediction, so we used synonyms to see how the model understood it. We changed the word "open" in the figure to unfold, which is the closest synonym for open. The outcome demonstrated that the likelihood of implausibility increased. We assume that the word "unfold" is close to a negative word, but it is actually a positive word in this context. Furthermore, we removed the word "open" from the sentence, and the model's prediction changed, increasing the likelihood of the story being implausible. This particularly shows that the model prediction is mainly dependent on one word instead of the whole sentence structure.

In our final analysis, the model also focuses

S/N	Human	Robot	Observation
1	Human: 1. Tom bought a new dustbin for the kitchen. 2. Tom threw a broken plate in the dustbin. 3. Tom got some soup from the fridge. 4. Tom put the soup in the microwave. 5. Tom turned on the microwave. (With two example prompts - Ann's story)	Robot: 1. Tom is a person. A person can buy a dustbin. A person can throw a broken plate in the dustbin. A person can get soup from a fridge. A person can put soup in a microwave. A person can turn on a microwave. 2. A dustbin can be bought by a person. A broken plate can be thrown into a dustbin. Soup can be retrieved from a fridge. Soup can be put in a microwave. A microwave can be turned on by a person. 3. Everything makes sense so this story is plausible.	Correct answer: However, there might be a bias to always validate the story as true
2	Human: 1. Tom bought a new dustbin for the kitchen. 2. Tom threw a broken plate in the dustbin. 3. Tom got some soup from the fridge. 4. Tom put the soup in the microwave. 5. Tom turned on the microwave. (With two example prompts - Ann's story)	Robot: 1. Tom is a person. A person can buy a dustbin. A person can throw a broken plate in the dustbin. A person can get some soup from the fridge. A person can put the soup in the microwave. A person can turn on the microwave. 2. A dustbin can be bought by a person. A broken plate can be thrown in a dustbin. 3. Soup can be stored in a fridge. 4. Soup can be put in a microwave. 5. A microwave can be turned on by a person. 6. Everything makes sense so this story is plausible.	Correct answer: However, there might be a bias to always validate the story as true

Figure 4: Predictions for Plausible Stories

on less important words in the story classification, such as an article (e.g. the). In general, the model is incapable of comprehending time series relationships between sentences. This can be seen in the first two Figures 3, where the two conflicting sentences, for example, "John open the book" and "John read the book," change position but the story predictions are plausible.

5.2 Results from Prompt Guided Dialog

Figure 4 and Figure 5 show a few predictions from GPT-3 for each story in our experiment. The tables below have been truncated to fit the constraints of this paper. Links to the full tables are provided in Appendix.

The overall results for GPT-3 on story classification using guided prompt dialog are summarized in the confusion matrix shown below in Figure 6.

The above results show that while the GPT-3 had 50 percent accuracy on the sample dataset used

S/N	Human	Robot	Observation
1	Human: 1. Tom bought a new dustbin for the kitchen. 2. Tom threw a broken plate in the dustbin. 3. Tom got some soup from the fridge. 4. Tom put the soup in the microwave. 5. Tom ate the cold soup. (With two example prompts - Ann's story)	Robot: 1. Tom is a person. A person can buy a dustbin. A person can throw a broken plate in a dustbin. A person can get soup from a fridge. A person can put soup in the microwave. A person can eat soup. 2. A dustbin can be bought by a person. A broken plate can be thrown in a dustbin. Soup can be taken from a fridge. Soup can be put in a microwave. Soup can be eaten by a person. 3. Everything makes sense so this story is plausible.	Wrong Accuracy Predicted Plausible when the story is actually Implausible (Why) Model ignored the physical state "cold" in sentence 5
2	Human: 1. Tom bought a new dustbin for the kitchen. 2. Tom threw a broken plate in the dustbin. 3. Tom unplugged the microwave. 4. Tom put the soup in the microwave. 5. Tom turned on the microwave. (With two example prompts - Ann's story)	Robot: 1. Tom is a person. A person can buy a dustbin. A person can throw a broken plate into the dustbin. A person can unplug a microwave. A person can put soup in the microwave. A person can turn on a microwave. 2. A dustbin can be bought by a person. A broken plate can be thrown into a dustbin. 3. A microwave can be unplugged by a person. 4. Soup can be put in a microwave. 5. A microwave can be turned on if it is plugged in. Everything makes sense so this story is plausible.	Wrong Accuracy While the model listed the assumptions correctly, it failed to draw the assumption from sentence 3 that will lead it to the correct answer.

Figure 5: Predictions for Implausible Stories

		Predicted	
		Plausible	Implausible
Actual	Plausible	5	0
	Implausible	5	0

Figure 6: Overall prompt guided classification results

for this analysis, the model was biased towards classifying every story as plausible. These results confirm that the task of story classification, even with the addition of assumptions, is non-trivial.

In cases of incorrect classification, further questions were added to the ongoing dialog to probe the reasoning of the GPT-3 and identify the flaws in the prediction process of the model. Figure 7 displays a few examples of the questions used to investigate model reasoning. Links to the unabridged version of the table can be found in the Appendix.

S/N	Human	Robot	Observation
1	Human: The story is actually implausible. Can you find out your mistake? (With two example prompts - Ann's story)	Robot: Yes. The story was implausible because it was impossible for Tom to eat cold soup after he put it in the microwave.	Model was able to backtrack after correction to derive the reasoning for the correct answer (involving physical state)
2	Human: The story is actually implausible. Can you find out your mistake?	Robot: I'm sorry, I don't understand that you mean. Could you please explain your question better?	Model was not able to self-correct in this case
2	Human: Is the microwave plugged in?	Robot: The story does not explicitly state if the microwave is plugged in. We can assume that the microwave is plugged in, since Tom was able to turn on the microwave.	The model over inferred leading it to believe the story is plausible
2	Human: However, there is a sentence in the story that states that Tom unplugged the microwave.	Robot: Yes, the sentence in the story states that Tom unplugged the microwave. However, the fact that he was able to turn on the microwave implies that it was plugged in before he unplugged it.	The model over inferred leading it to believe the story is plausible
2	Human: Since the story mentioned that Tom unplugged the microwave, we cannot state that the microwave was plugged back in unless the story explicitly states that it was plugged back in. In this case, is the story plausible or not? (With two example prompts - Ann's story)	Robot: The story is not plausible, since Tom was able to turn on the microwave even though it was unplugged. The reason that the microwave was plugged in, which contradicts the sentence in the story that states that Tom unplugged the microwave.	When the model drops the misleading inference, it is able to come to the right conclusion.

Figure 7: Questions for Probing Model Reasoning

6 Discussion

The LIME experiment's findings provided some crucial insights, particularly regarding how the large language model BERT acquired the information. We used the explainability tool from the viewpoint of a model designer in addition to met-

rics like accuracy to better understand how the data was interpreted. LIME's fundamental idea might seem straightforward, but there are a few possible cons.

In the current implementation, only linear models are used to approximate local behavior. When focusing on a very small area close to the data sample, this assumption is mostly accurate. However, it's possible that by enlarging this region, a linear model won't be strong enough to account for the behavior of the original model. When using datasets that call for complicated, unintelligible models, nonlinearity in local regions occurs. A significant disadvantage is that LIME cannot be used in these circumstances.

Second, the types of changes that must be made to the data in order to obtain proper explanations are typically use-case specific. Simple perturbations are frequently insufficient. The modifications should ideally be driven by the variation observed in the dataset. Manually steering the instabilities on the other hand is probably not a good idea because it will almost certainly introduce bias into the model explanations.

Aside from LIME's limitations, the BERT model we fine-tuned has low accuracy, which may have an impact on outcome analysis. We tried a variety of hyperparameters when fine-tuning, but the accuracy remained constant. First, we hypothesize that the low accuracy is due to the TRIP dataset's preparation for multiple-tired annotation, which made preprocessing the story classification part and balancing the dataset's plausible and implausible stories difficult.

One hypothesis that we had before conducting the prompt-guided dialog experiment was that few-shot prompting of GPT-3 with assumptions as additional information would yield a high story classification accuracy. However, this was not the case. While the prompt-guided dialog approach did not yield solid empirical results, its strength lies in allowing researchers to trace the model's decision process and identify if language models use logical reasoning steps to make predictions. In this case, GPT-3 was given a few examples that extracted assumptions from the given stories and classified the stories based on the extracted assumptions. The responses returned by the model clearly show that the model was able to learn the task of extracting assumptions, however, this did not automatically translate into high performance in story classifica-

tion. In view of this, one could argue that not every setup for multitask learning will lead to optimal model learning of two or more tasks. For humans, improving knowledge of a related task often leads to better performance in another task. However, the same might not be accurate for high-performing models that might use other methods of reasoning unfamiliar to humans. Unfortunately, this theory is hard to investigate since most SOTA language models are black boxes.

Analysis of the model’s responses to probing questions raised awareness about several possible reasons why the model was underperforming in the story classification task. Behaviors such as rearranging the sentences to fit common sense before classifying the story, constructing more assumptions than necessary to fill in the inconsistency, and checking if the actions mentioned are possible events within the scope of a single sentence alone might be pointers to areas where the model architecture or dataset can be improved for better performance in the task. Since this experiment lacks statistical power, the way forward would be to test the model on a larger dataset using model configurations designed to investigate and address the issues raised by this experiment.

7 Conclusion and Future Work

In this work, we attempted to understand model behavior in addition to metrics such as accuracy. We look deeper into how the model interprets the given data from the TRIP dataset. To unpack the models’ black box behavior, we proposed an explainability tool LIME, and a prompt-guided dialog approach with GPT-3. We believe that these two methods will assist model developers in better understanding how human mental models and model learning patterns match. This can also help to improve the interpretability of large language models as well as identify flaws in trained models and datasets.

For our future work, we want to investigate several different potential trajectories. Despite the fact that we used LIME, we want to try out other tools that are explainable and effective with large language models in order to better understand the internal behavior of the model. We also aim to improve the accuracy of the current model and check whether the LIME interpretation has changed. In addition to improving model accuracy, we also want to understand why the models from the previous papers and other large language models don’t

understand common sense. One analysis direction we plan to attempt in upcoming works is how time series data are interpreted and comprehended in complex language models. Another direction would involve the analysis of the natural language model explanations generated by SOTA language models on the entire TRIP dataset and similar datasets like (Gordon et al., 2011).

References

- Michael Ahn, Anthony Brohan, Noah Brown, Yevgen Chebotar, Omar Cortes, Byron David, Chelsea Finn, Chuyuan Fu, Keerthana Gopalakrishnan, Karol Hausman, Alex Herzog, Daniel Ho, Jasmine Hsu, Julian Ibarz, Brian Ichter, Alex Irpan, Eric Jang, Rosario Jauregui Ruano, Kyle Jeffrey, Sally Jesmonth, Nikhil J Joshi, Ryan Julian, Dmitry Kalashnikov, Yuheng Kuang, Kuang-Huei Lee, Sergey Levine, Yao Lu, Linda Luu, Carolina Parada, Peter Pastor, Jornell Quiambao, Kanishka Rao, Jarek Rettinghouse, Diego Reyes, Pierre Sermanet, Nicolas Sievers, Clayton Tan, Alexander Toshev, Vincent Vanhoucke, Fei Xia, Ted Xiao, Peng Xu, Sichun Xu, Mengyuan Yan, and Andy Zeng. 2022. [Do as i can, not as i say: Grounding language in robotic affordances](#).
- Yonatan Bisk, Rowan Zellers, Ronan Le Bras, Jianfeng Gao, and Yejin Choi. 2019. [PIQA: reasoning about physical commonsense in natural language](#). *CoRR*, abs/1911.11641.
- Antoine Bosselut, Hannah Rashkin, Maarten Sap, Chaitanya Malaviya, Asli Celikyilmaz, and Yejin Choi. 2019. [COMET: commonsense transformers for automatic knowledge graph construction](#). *CoRR*, abs/1906.05317.
- 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. 2020. [Language models are few-shot learners](#).
- Oana-Maria Camburu, Brendan Shillingford, Pasquale Minervini, Thomas Lukasiewicz, and Phil Blunsom. 2020. [Make up your mind! adversarial generation of inconsistent natural language explanations](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4157–4165, Online. Association for Computational Linguistics.
- Mayank Chhipa, Hrushikesh Mahesh Vazurkar, Abhijeet Kumar, and Mridul Mishra. 2020. [Comparative study of language models on cross-domain](#)

- [data with model agnostic explainability](#). *CoRR*, abs/2009.04095.
- Ernest Davis and Gary Marcus. 2015. [Commonsense reasoning and commonsense knowledge in artificial intelligence](#). *Commun. ACM*, 58(9):92–103.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.
- Andrew Gordon, Cosmin Bejan, and Kenji Sagae. 2011. Commonsense causal reasoning using millions of personal stories. volume 2.
- Mayank Kejriwal and Ke Shen. 2020. [Do fine-tuned commonsense language models really generalize?](#) *CoRR*, abs/2011.09159.
- Christoph Leiter, Piyawat Lertvittayakumjorn, Marina Fomicheva, Wei Zhao, Yang Gao, and Steffen Eger. 2022. [Towards explainable evaluation metrics for natural language generation](#).
- Kavi Mahesh, Sergei Nirenburg, Jim Cowie, and David Farwell. 1996. An assessment of cyc for natural language processing.
- Bhargavi Paranjape, Julian Michael, Marjan Ghazvininejad, Luke Zettlemoyer, and Hananeh Hajishirzi. 2021. [Prompting contrastive explanations for commonsense reasoning tasks](#). *CoRR*, abs/2106.06823.
- Fabio Petroni, Tim Rocktäschel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, Alexander H. Miller, and Sebastian Riedel. 2019. [Language models as knowledge bases?](#)
- Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. ["why should i trust you?": Explaining the predictions of any classifier](#). In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '16*, page 1135–1144, New York, NY, USA. Association for Computing Machinery.
- Maarten Sap, Ronan LeBras, Emily Allaway, Chandra Bhagavatula, Nicholas Lourie, Hannah Rashkin, Brendan Roof, Noah A. Smith, and Yejin Choi. 2018. [ATOMIC: an atlas of machine commonsense for if-then reasoning](#). *CoRR*, abs/1811.00146.
- Vered Shwartz, Peter West, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. 2020. [Unsupervised commonsense question answering with self-talk](#).
- Robyn Speer, Joshua Chin, and Catherine Havasi. 2016. [Conceptnet 5.5: An open multilingual graph of general knowledge](#). *CoRR*, abs/1612.03975.
- Shane Storks, Qiaozi Gao, Yichi Zhang, and Joyce Chai. 2021a. [Tiered reasoning for intuitive physics: Toward verifiable commonsense language understanding](#). *CoRR*, abs/2109.04947.
- Shane Storks, Qiaozi Gao, Yichi Zhang, and Joyce Chai. 2021b. [Tiered reasoning for intuitive physics: Toward verifiable commonsense language understanding](#).
- Mateusz Szczepański, Marek Pawlicki, Rafał Kozik, and Michał Choraś. 2021. [New explainability method for bert-based model in fake news detection](#). *Scientific Reports*, 11.
- Trieu H. Trinh and Quoc V. Le. 2019. [Do language models have common sense?](#)
- Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Ed H. Chi, Quoc Le, and Denny Zhou. 2022. [Chain of thought prompting elicits reasoning in large language models](#). *CoRR*, abs/2201.11903.

A Example Appendix

A.1 Plausible Stories

[Link to Plausible Story Predictions](#)

[Link to Plausible Story Investigation Questions](#)

A.2 Implausible Stories

[Link to Implausible Story Predictions](#)

[Link to Implausible Story Investigation Questions](#)