Improve Model Performance on TRIP Dataset

Jiaqi Xu
University of Michigan
xjiaqi@umich.edu

Yukun Yang
University of Michigan
yyukun@umich.edu

Tianao Chen
University of Michigan
tianaoch@umich.edu

Chuying Han
University of Michigan
chuyingh@umich.edu

Abstract

Understanding physical common-sense reasoning is critical for advancing artificial intelligence systems. The Tiered Reasoning for Intuitive Physics (TRIP) dataset provides a comprehensive framework for evaluating AI models on tasks requiring nuanced reasoning about plausibility, conflict detection, and physical state transitions in narrative contexts. Despite its importance, the dataset's small size limits model generalization. In this study, we address these limitations through five approaches: leveraging advanced pre-trained models like LLaMA-2-7B, implementing transfer learning with external datasets such as PIQA and HellaSwag, applying data augmentation techniques-including object abstraction, synonym replacement, external dataset, and knowledge graph integration with ConceptNet and COMET-utilizing meta-learning, and prompting large language models for in-context learning. Our experiments reveal improvements in accuracy, consistency, and verifiability, demonstrating the potential of these strategies to enhance common-sense reasoning capabilities in AI. These findings not only advance the state-of-the-art for TRIP tasks but also provide a robust foundation for addressing broader challenges in explainable AI and physical common-sense reasoning.

1 Introduction

Understanding physical common-sense reasoning is a critical task in advancing artificial intelligence systems. Unlike tasks that involve basic pattern recognition or simple text understanding, reasoning about intuitive physics requires models to capture complex relationships between entities, actions, and their consequences. This challenge has motivated the development of large-scale pre-trained language models such as BERT (Devlin et al., 2019), RoBERTa (Liu et al., 2019), which have shown remarkable success in various natural language processing tasks. Despite their strengths, these models often struggle with tasks requiring deeper reasoning and common-sense understanding (Bender and Koller, 2020).

The Tiered Reasoning for Intuitive Physics (TRIP) dataset addresses this challenge by evaluating models on their ability to differentiate plausible and implausible stories, identify conflicting sentences in implausible narratives, and track the state transitions of entities that cause these conflicts (Storks et al., 2021). These tasks reflect a multi-faceted reasoning framework, making TRIP a benchmark dataset for advancing explainable AI in common-sense reasoning. However, its relatively small size poses challenges for training robust models, similar limitations observed in datasets like ROCStories (Mostafazadeh et al., 2016) and SPARTQA (Mirzaee et al., 2021).

To overcome these limitations, integrating external datasets and leveraging advanced pre-trained models have emerged as promising directions. For instance, datasets such as PIQA (Bisk et al., 2019) and ROCStories (Mostafazadeh et al., 2016) provide narrative coherence and physical interaction reasoning patterns, while SPARTQA (Mirzaee et al., 2021) emphasizes spatial reasoning and GITA (Pensa et al., 2024)introduces linguistic diversity by expanding tasks to non-English languages. These datasets complement TRIP by introducing a broader range of reasoning scenarios, which can be leveraged through transfer learning to improve the robustness and generalization of models (Raffel et al., 2023).

In this project, we address the limitations of existing approaches by employing five core strategies. First, we apply advanced pre-trained models to enhance contextual understanding and reasoning capabilities, enabling a deeper comprehension of physical common-sense tasks. Second, we utilize transfer learning by pre-training models on related datasets before fine-tuning them on TRIP tasks, achieving better task-specific performance.

Third, we implement diverse data augmentation techniques, including knowledge graph integration from, as well as in-domain methods such as object abstraction and synonym replacement, to enrich the training data. Fourth, we employ in-context learning by prompting large language models, including GPT-3, Mistral, and Llama, to address reasoning tasks with few-shot adaptability. Finally, we incorporate meta-learning, which utilizes a support-query framework to generalize across tasks, providing a scalable approach for improving task adaptability and robustness.

These strategies are designed to improve the accuracy, consistency, and verifiability of TRIP tasks, while establishing a more comprehensive framework for tackling physical common-sense reasoning challenges.

2 Related Work

TRIP. The Tiered Reasoning for Intuitive Physics (TRIP) (Storks et al., 2021) is a benchmark for physical common-sense reasoning that provides traces of reasoning for an end task of plausibility prediction. The dataset consists of human-authored stories, such as those in Figure 1, describing sequences of concrete physical actions. Given two stories composed of individually plausible sentences and only differing by one sentence (i.e., Sentence 5), the proposed task is to determine which story is more plausible. To understand stories like these and make such a prediction, one must have knowledge of verb causality and precondition, and rules of intuitive physics. Plausible stories were crowd-sourced from Amazon Mechanical Turk. To convert each story into several implausible stories, separate workers were hired to each write a new sentence to replace a sentence in the original story, such that the new story after replacement is no longer realistic in the physical world. To ensure quality, these workers flagged stories which were incoherent or did not describe realistic actions. TRIP eliminated those stories and performed a manual round of validation to remove any remaining bad stories and correct typos.

A baseline model for the TRIP task was also proposed by (Storks et al., 2021). As illustrated in Figure 2, the model consists of four components. It starts with a contextual embedding layer which takes an sentence, the name of entity, and an entity-first formulation as input, and outputs a contextualized numerical representation. This module is initially implemented with a BERT-like pre-trained language model. Following this, the Precondition and Effect Classifiers aims to classify preconditions and effects into 20 predefined physical attributes. A conflict detector is then used to identify conflicts in the entity's physical state and locate a pair of conflicting sentences. Finally, the model includes a story choice classifier which determines the more plausible story.

Data Augmentation. In recent years, various studies have been working to enrich datasets that can help enhance reasoning models' capability to understand the rich interactions between agents, locations, and events. (Asai and Hajishirzi, 2020) utilized logical and linguistic knowledge to augment labeled training data, achieving large improvements over question answering and reading comprehension tasks. (Jiang et al., 2023) experimented with both in-domain and out-of-domain data augmentation strategies on procedural understanding tasks, demonstrating that these strategies can lead to higher accuracy and help models generalize better to unseen stories. Data augmentation has been extensively studied as an effective technique for improving model robustness. (Wei and Zou, 2019) proposed Easy Data Augmentation (EDA), which uses techniques such as synonym replacement, random insertion, and random deletion to generate diverse training samples. These methods have been shown to significantly improve performance on small datasets by reducing over-fitting and enhancing model generalization. From another perspective, there are also ongoing efforts to generate new datasets. By creating a corpus of five-sentence common-sense stories, (Mostafazadeh et al., 2016) provided a new evaluation framework, ROCStories, for language understanding tasks. In addition, (Ilievski et al., 2021) combined common-sense axioms with common-sense knowledge graphs to generate synthetic stories, enriching the story corpus for training reasoning models. In the work of (Pensa et al., 2024), a tiered Italian dataset called Graded Italian Annotated dataset (GITA) was created to evaluate large language models' different levels of common-sense understanding in Italian.

Knowledge Graph for Commonsense Reasoning. While BERT-like pre-trained language models excel in many tasks, they often focus on embedding the content of sentences and lack common-sense knowledge required for reasoning tasks. The integration of structured knowledge into language models has been shown to enhance their performance on

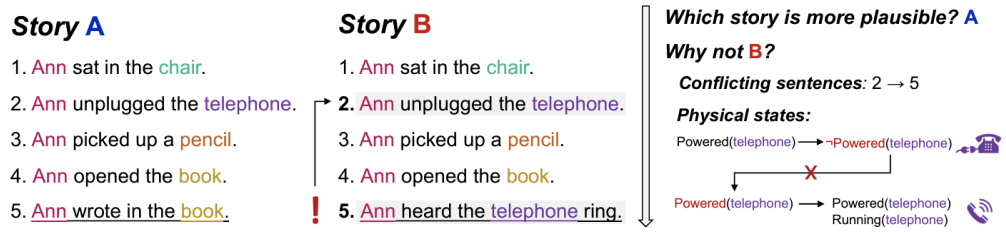


Figure 1: Story pair from TRIP, along with the tiers of annotation available to represent the reasoning process.

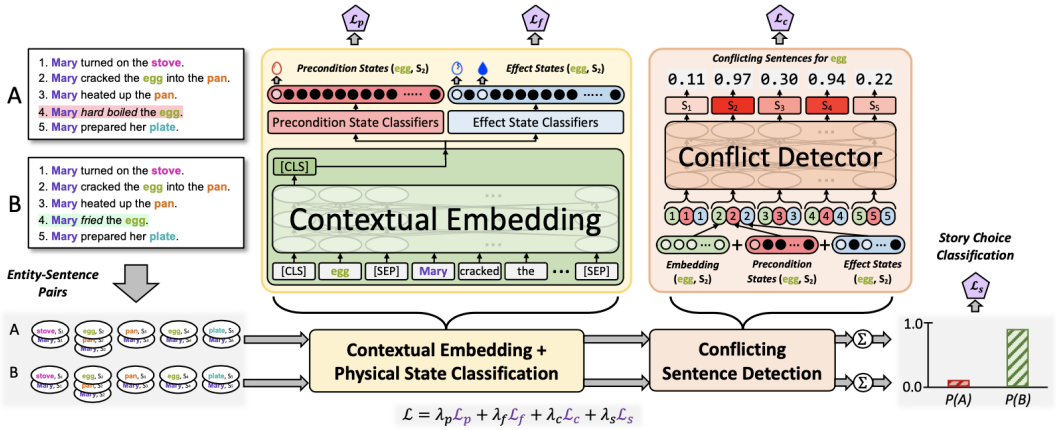


Figure 2: Proposed tiered reasoning system in (Storks et al., 2021).

tasks requiring common-sense reasoning. (Zhang et al., 2019) introduced ERNIE, which incorporates informative entities from knowledge graphs like Wikidata into pre-trained language models, achieving state-of-the-art results on several benchmarks. Inspired by this, we incorporate ConceptNet and COMET to enrich the contextual understanding of the TRIP dataset. ConceptNet (Speer et al., 2017) is a knowledge graph that connects words and phrases of natural language with labeled edges. By providing explicit connections between entities such as UsedFor, LocationOf, EventOf, and SubeventOf, ConceptNet enables machines to reason more effectively about the everyday contexts. Fine-tuned on ConceptNet and ATOMIC (Sap et al., 2019), COMmonsEnse Transformers (COMET) (Bosselut et al., 2019) has been developed to automatically generate diverse relationship triples beyond what is explicitly stored in static graphs, including attributes of entities, consequences of events on subjects and objects, intentions behind actions, preconditions needed for actions, and so on. This helps capture both implicit and explicit information, thus advancing common-sense understanding in natural language processing tasks.

In-context Learning with LLMs. The concept of in-context learning (ICL) was first introduced by (Brown, 2020). Large language models (LLMs) such as LLaMA2 (Touvron et al., 2023) and GPT-4 (OpenAI, 2023) have exhibited amazing performance across a wide range of tasks, including machine translation, text classification, and question answering, when prompted with a minimal set of task-oriented examples, all without in-domain gradient-based training. The exploration of different in-context learning and sequential prompting approaches (Wang et al., 2022; Long, 2023; Yao et al., 2023) has further enhanced the reasoning capabilities of LLMs across various tasks. By decomposing a high-level task into many low-level sub-tasks which the models can solve easily, models' understanding in higher-level tasks can be enhanced. In the work by (Zhang et al., 2023), various techniques such as unstructured in-context learning, chain-of-thought in-context learning, and heuristic-analytic in-context learning, have been investigated to refine the generation of low-level common-sense knowledge from LLMs in complex cases where lower-level sub-tasks are hard to solve due to the requirement of retrieving and incorporating knowledge beyond the text.

Meta-Learning. Meta-learning, often referred to

as "learning to learn," has emerged as a promising approach in tasks with limited labeled data. Unlike traditional machine learning approaches that train models on a fixed dataset, meta-learning optimizes a model's ability to adapt quickly to new tasks. This is typically achieved by training on a variety of related tasks, enabling the model to extract shared knowledge across tasks and generalize effectively to unseen tasks.

One prominent framework in meta-learning is Model-Agnostic Meta-Learning (MAML) (Finn et al., 2017), which aims to learn a set of initial model parameters that can be fine-tuned with minimal updates for a specific task. MAML has shown success in areas such as few-shot classification (Snell et al., 2017) and reinforcement learning.

In the context of this project, we adopt a meta-learning approach tailored to our specific dataset. The method involves constructing support and query sets for training and evaluation, enabling the model to learn from limited data while achieving strong performance in generalization tasks. This approach builds upon the insights from prior meta-learning works while being adapted to address the unique challenges of our experimental setting.

3 Dataset

TRIP Dataset. The TRIP dataset provides a strong foundation for evaluating physical plausibility and consistency. It includes dense annotations for plausibility classification, conflict detection, and physical state prediction. TRIP's tiered tasks make it a comprehensive and challenging benchmark.

However, the limited size of TRIP dataset poses challenges for model generalization and performance. To address these issues, we incorporate external datasets such as ROCStories, PIQA, and SPARTQA. These datasets complement TRIP by offering diverse linguistic patterns, causal reasoning structures, and physical interaction scenarios. By integrating these datasets, we aim to expand training diversity, improve model robustness, and facilitate transfer learning for better performance on TRIP's complex multi-task framework.

Propara Dataset. The Propara dataset is designed by (Dalvi et al., 2018) to track the state changes of entities in procedural text. It focuses on tasks like identifying when and where an entity changes state or location. Though simpler than TRIP, Propara provides a structured approach to entity state tracking, which is useful for pre-training models to understand sequential transformations.

PIQA Dataset. The Physical Interaction Question Answering (PIQA) dataset was introduced by (Bisk et al., 2019) to assess physical reasoning in NLP models. It contains multiple-choice questions about everyday physical interactions, requiring models to select the most plausible answer. This dataset highlights the importance of understanding physical affordances and object properties, which aligns with TRIP's physical common-sense tasks.

HellaSwag Dataset. The HellaSwag dataset was introduced by (Zellers et al., 2019) to evaluate the ability of NLP models to perform common-sense reasoning, particularly for situations that involve narrative completion. It contains multiple-choice questions where the task is to select the most plausible continuation of a given context. HellaSwag emphasizes the importance of common-sense knowledge and understanding of everyday scenarios, highlighting challenges where models can fail due to superficial reasoning.

SPARTQA Dataset. The SPATial Reasoning Question Answering (SPARTQA) dataset was proposed by (Mirzaee et al., 2021) to model realistic spatial reasoning phenomena. It includes scenarios requiring models to infer spatial relationships between entities. SPARTQA's structured annotations provide valuable insights for spatial conflict detection, complementing TRIP's focus on reasoning about object positions and movements.

ROCStories Dataset. ROCStories is a corpus of short stories introduced by (Mostafazadeh et al., 2016). It contains a variety of common-sense causal and temporal relations between everyday events. Each of the 3,742 story pairs consists of a four-sentence-long body and two candidate endings, with only one of the two endings plausible given the previous four sentences.

GITA Dataset. Graded Italian Annotated (GITA) dataset is developed by (Pensa et al., 2024) to analyze physical common-sense reasoning in large language models for Italian. It consists of 355 story combinations, where each plausible story is paired with two implausible stories: one containing conflicts between sentences and the other has inverse order issues. Similar to TRIP, GITA also provides annotations for the state of each entity in each sentence.

4 Approaches

4.1 Advanced Pre-trained Models

The models used in the original paper are all BERT-like models. To further examine how non-BERT models perform on the three tasks of the TRIP dataset, we selected Llama-2-7B for training and testing. Llama-2-7B is based on the Transformer architecture and leverages a corpus that supports common-sense reasoning during pre-training. This design aims to improve its ability to interpret contextual information and logical relationships.

By conducting experiments on the TRIP dataset, we seek to determine whether this architecture and training approach can enhance performance on the three sub-tasks, including the final story classification task. To ensure the comparability of results, we maintained the same hyperparameters and experimental conditions used for BERT-like models. This consistency allows us to more clearly compare the results across models in terms of common sense inference and story comprehension, thereby examining how specific architectural features and training strategies influence task outcomes.

4.2 Fine-tuning

Fine-tuning can improve a model's generalization ability and prediction accuracy by allowing it to more effectively adapt to the feature distribution and semantic requirements of a specific task. Rather than training from scratch, fine-tuning also reduces time and resource consumption, since the model already possesses basic feature representation and language understanding capabilities. Following these principles, we took a BERT model trained on the PIQA dataset and a RoBERTa model trained on the HellaSwag dataset, and fine-tuned both on the TRIP dataset.

For the fine-tuning process, we employed a systematic approach across multiple training phases. First, we initialized both models with their respective pre-trained weights from PIQA and HellaSwag datasets, which provided foundational common-sense reasoning capabilities. The PIQA dataset, focusing on physical common-sense knowledge, and HellaSwag, emphasizing situation understanding, were specifically chosen for their alignment with TRIP's reasoning requirements. During fine-tuning, we maintained consistent hyperparameter settings across all experiments to ensure fair comparison: a learning rate of 1e-5, batch size of 1, and maximum sequence length of 16. And we tested the models' performance on all three tasks, including the final story classification task.

4.3 Data Augmentation

The TRIP dataset is relatively small in scale due to the laborious story creation and attribute labeling. This could pose challenges for training models that can generalize well. In fact, (Storks et al., 2021) mentioned that while the losses for state classification loss, conflict detection loss, and story choice on the training dataset decreased consistently during the model training process, these losses on the validation dataset started to oscillate and even increase at an early stage. This behavior indicates model's over-fitting on low-level tasks. Therefore, we plan to use data augmentation techniques to mitigate this problem.

4.3.1 In-Domain Augmentation

We apply two in-domain data augmentation methods to the original TRIP dataset: Object Abstraction and Synonym Replacement.

Object Abstraction. We replace all non-human objects in the stories with their direct hypernyms in WordNet. This process abstracts specific objects into more general categories, encouraging the model to focus on the relationships and conflicts within the stories rather than the detailed entities. Below are some examples.

pizza, soup, salad, ... $\rightarrow$ *dish*
box, dustbin, ... $\rightarrow$ *container*
door, ... $\rightarrow$ *movable barrier*

To achieve this, we first retrieve the synset for each object that appears in the story dataset from WordNet, and then extract its direct hypernym to construct an abstraction dictionary. This dictionary is subsequently applied to create a new object-abstracted dataset based on the original TRIP dataset.

Synonym Replacement. We replace all adjectives and adverbs in the sentences with their synonyms, where available. For verbs, we identify synonyms for their present tense forms, convert these synonyms into past tense, and then substitute them into the sentence.

4.3.2 External Dataset Augmentation

We incorporate a Graded ITalian Annotated (GITA) dataset (Pensa et al., 2024) to expand the training dataset in TRIP. The GITA dataset consists of 355 story combinations, where each plausible

story is paired with two implausible stories. Similar to TRIP, GITA also provides annotations for the state of each object in each sentence. However, the dataset is entirely written in Italian.

To utilize GITA for model training, we apply a neural machine translation model to translate it from Italian to English. Subsequently, plausible stories are paired with their corresponding conflicting implausible counterparts. We then manually fill in any missing information, such as the conflicting sentence and conflicting pair annotations, and reformat the dataset to align with TRIP's structure.

4.3.3 Knowledge Graph Integration

In order to enhance the reasoning capabilities of the model, we integrate knowledge from ConceptNet, a well-known knowledge graph, and COMET, a generative common-sense model, into the original TRIP dataset.

Specifically, for each entity in the sentence, we first query ConceptNet API to retrieve high-weighted relationships. This generates triples that can describe the attributes of entities such as 'CapableOf', 'AtLocation', and 'RelatedTo'. To ensure the quality of the information, the results are filtered to include only those with a weight higher than a certain threshold.

We then use COMET to generate additional relationships for each entity, including 'xAttr' (attributes), 'xEffect' (consequences for the subject), and 'oEffect' (consequences for the object). For example, for the entity 'microwave', xAttr = 'hot', oEffect = 'heat up food'.

Eventually, these relationships from ConceptNet and COMET are mapped to their relevant entities and are transformed into contextual descriptions that are appended to the original sentences.

By combining these descriptions with the original coherent stories, we aim to help the model better capture the functional roles of entities and understand the underlying causalities in the stories.

4.4 In-context Learning with LLMs

We follow the idea of in-context learning with chain of thought (ICL-CoT) as proposed in (Zhang et al., 2023) to generate prompts for large language models. Unlike (Zhang et al., 2023), which prompts InstructGPT and Llama-65B, we select free, smaller models with fewer parameters due to resource restrictions. Specifically, we focus on GPT-3, Mistral-7B, and Llama-3.1-8B.

Since the TRIP task requires both high-level predictions and justifications based on low-level physical common-sense evidence, the task is divided into four sub-tasks. In the first step, we prompt the model with "Let's think step by step about which story is more plausible.", asking it to identify the more plausible story. For story pairs where the model correctly predicts the plausible story, we move on to the second step, where we ask the model to justify its decision by using the prompt "Let's think step by step about which sentences are conflicting in one story." If the model provides the conflicting sentence pairs correctly, we ask for its further predictions of action precondition and action effect prediction. In this stage, we prompt the model with "Let's think step by step about which physical states are conflicting in two sentences in one story. In Story A, ... In Story B, ... Therefore: After, what is the state of [entity]? Before, what was the state of [entity]?" The results from each stage are extracted and stored for final evaluation.

This multi-stage prompting strategy enables the large language models to reason through both high-level tasks, such as plausibility identification, and low-level tasks, such as conflict detection and physical state analysis, enhancing its understanding of the contexts.

4.5 Meta-Learning

In addition to the baseline and pre-existing approaches, we implemented a **Meta-Learning Approach** using BERT for Sequence Classification. This method involves the use of *meta-tasks*, which are constructed from the dataset to train the model to generalize better on unseen tasks.

The dataset was partitioned into multiple tasks, each comprising a *support set* and a *query set*. The support set is used to train the model on task-specific samples, while the query set evaluates its generalization within the same task. Tasks were formed by grouping related sentences into logical contexts and labels (e.g., plausible vs. implausible scenarios).

The BERT-base model was fine-tuned as the backbone for learning task representations. For each task, the model processes support inputs to compute task-specific gradients, which are then used for query predictions.

A gradient-based meta-learning strategy was employed, iteratively optimizing task-specific loss functions over multiple epochs. The learning rate and batch sizes were tuned for stable training dy-

namics.

5 Evaluation

We assess the model's coherent reasoning ability from three perspectives: Accuracy, Consistency, and Verifiability.

Accuracy. This metric measures the proportion of examples for which the model identifies plausible stories correctly.

Consistency. This metric evaluates the proportion of examples where the model not only correctly distinguishes between plausible and implausible stories, but also accurately identifies the conflicting sentence pairs in the implausible stories.

Verifiability. This metric tests the proportion of examples where the model correctly identifies not only the plausible stories and the conflicting sentence pairs in the implausible stories, but also the underlying physical states that contribute to the conflicts.

Verifiability and Consistency are much stricter than Accuracy, and can better reflect the model's understanding of causalities in the stories and reasoning ability on low-level tasks. Given the poor performance of the baseline model (Storks et al., 2021) on these two metrics, we aim to improve them through adjustments to the model architecture and training process.

6 Results

6.1 Advanced Pre-trained Models and Fine-tuning

In our experiments on transfer learning and non-BERT-like models, we employed the Llama-2-7B model, as well as the BERT model pre-trained on the PIQA dataset and the RoBERTa model pre-trained on the HellaSwag dataset. During training, we utilized a combined loss and the Omit Story Choice Loss, where the former demonstrated excellent performance in prediction accuracy, while the latter excelled in ensuring result consistency and verifiability. The corresponding experimental results are presented in Table 1 and Table 2, covering three configurations: Llama-2-7B, original+BERT-PIQA, and original+RoBERTa-HellaSwag.

Our findings show that despite lacking task-specific pre-training, the Llama-2-7B model achieved exceptional performance across all metrics. Specifically, when trained with combined loss, Llama-2-7B showed improvements of 2.5% and 2.3% in verifiability and consistency respectively compared to the baseline, suggesting that large language models possess inherent capabilities for complex reasoning tasks. Regarding transfer learning approaches, BERT pre-trained on PIQA exhibited a modest accuracy improvement of 3% over the baseline when using combined loss. However, the model demonstrated a trade-off between accuracy and verifiability, with consistency scores declining by 0.7% compared to the baseline. Similarly, RoBERTa pre-trained on HellaSwag achieved a 4% accuracy gain but experienced a comparable decline in consistency metrics. Upon omitting story choice loss, we observed that overall accuracy decreased slightly across all models, but consistency and verifiability metrics showed significant improvements. This suggests that the omitting story choice loss effectively guides models toward more reliable and verifiable reasoning patterns.

Model	Accuracy (%)	Consistency (%)	Verifiability (%)
Baseline			
BERT	78.3	2.8	0.0
RoBERTa	75.2	6.8	0.9
Advanced Models			
Llama-2-7B	62.1	9.3	3.1
Pre-trained Models + Fine-tuning			
BERT-PIQA	81.3	2.1	0.0
RoBERTa-HS	79.2	6.3	0.1

Table 1: Experimental evaluation on the test dataset using All Losses (Advanced models and fine-tuning)

Model	Accuracy (%)	Consistency (%)	Verifiability (%)
Baseline			
BERT	73.9	28.0	9.0
RoBERTa	73.6	22.4	10.6
Pre-trained Models + Fine-tuning			
BERT-PIQA	76.5	17.7	5.1
RoBERTa-HS	77.5	16.5	3.1

Table 2: Experimental evaluation on the test dataset when omitting Story Choice Loss (Advanced models and fine-tuning)

6.2 Data Augmentation

In our data augmentation experiments, we focus on BERT and RoBERTa models. We use both all losses and the configuration where the story choice loss is omitted during training process, as these could potentially yield the best verifiability as indicated in (Storks et al., 2021). We create five augmented datasets by applying the following

techniques: (1) Object Abstraction, (2) Synonym Replacement, (3) both Object Abstraction and Synonym Replacement jointly, (4) adding the GITA dataset, and (5) Knowledge Graph Integration. Experiments are conducted on 6 training dataset combinations: (1) original training dataset + object abstracted dataset, (2) original training dataset + synonym replaced dataset, (3) original training dataset + object abstracted dataset + synonym replaced dataset, (4) original training dataset + both object abstracted and synonym replaced dataset, (5) original training dataset + GITA dataset, and (6) original training dataset + knowledge graph integrated dataset.

Model	Accuracy (%)	Consistency (%)	Verifiability (%)
Baseline			
BERT	78.3	2.8	0.0
RoBERTa	75.2	6.8	0.9
Data Augmentation			
original + object abstracted dataset			
BERT	71.5	6.6	2.0
RoBERTa	78.4	6.0	2.0
original + synonym replaced dataset			
BERT	76.9	5.7	2.0
RoBERTa	77.8	3.4	1.4
original + GITA dataset			
BERT	73.5	4.3	0.6
RoBERTa	75.8	4.6	1.2

Table 3: Experimental evaluation on the test dataset using All Losses (Data augmentation)

The experimental metrics on the test dataset for the BERT and RoBERTa models trained on various augmented datasets using all losses are presented in Table 3. Under this circumstance, data augmentation techniques help BERT models achieve an increase in consistency from 2.8% to 6.6% and an improvement in verifiability from 0.0% to 2.0%, although at the cost of accuracy. Meanwhile, RoBERTa models exhibit slighter improvements in verifiability but more gains in accuracy, both at the expense of consistency.

Table 4 shows the experimental metrics on the test dataset for the BERT and RoBERTa models trained on various augmented datasets when the story choice loss is omitted. The results indicate that object abstraction and synonym replacement can slightly improve the verifiability of BERT models with little impact on the overall accuracy, while incorporating additional GITA dataset and knowledge graph lead to worse performance compared

Model	Accuracy (%)	Consistency (%)	Verifiability (%)
Baseline			
BERT	73.9	28.0	9.0
RoBERTa	73.6	22.4	10.6
Data Augmentation			
original + object abstracted dataset			
BERT	73.8	22.8	11.7
RoBERTa	76.1	23.9	12.3
original + synonym replaced dataset			
BERT	74.9	24.2	10.8
RoBERTa	71.2	23.1	11.4
original + object abstracted + synonym replaced dataset			
BERT	73.2	22.8	9.0
RoBERTa	72.9	25.9	11.4
original + object abstracted & synonym replaced dataset			
BERT	72.4	23.1	11.4
RoBERTa	77.8	24.2	12.5
original + GITA dataset			
BERT	77.8	20.8	8.3
RoBERTa	77.8	27.1	11.1
original + knowledge graph integrated dataset			
BERT	72.4	23.1	8.6
RoBERTa	75.5	27.4	13.4

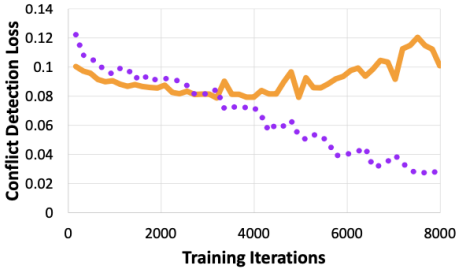
Table 4: Experimental evaluation on the test dataset when omitting Story Choice Loss (Data augmentation)

to the baseline. In contrast, all the data augmentation techniques work for RoBERTa models. The BERT model achieves its best verifiability of 11.7% when trained on the original dataset combined with the object-abstracted dataset. Meanwhile, the RoBERTa model reaches the highest verifiability of 13.4% with the incorporation of information generated from knowledge graph.

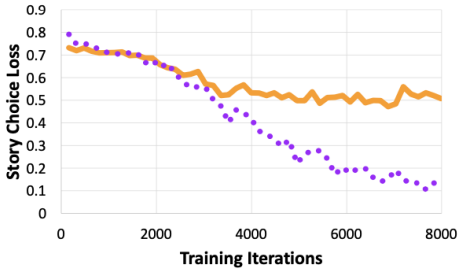
Figure 3 compares the losses of the baseline model from (Storks et al., 2021) and our model trained on the augmented dataset. Similar to the baseline model, the model trained on the augmented dataset still suffers from over-fitting issue, with conflict detection loss in the training dataset decreasing consistently while that in the validation dataset starting to rise at a very early stage. The conflict detection loss in the validation dataset exhibits the same pattern across all our data augmentation techniques. This suggests that although slight improvements in verifiability have been achieved through data augmentation, the over-fitting issue remains unsolved.

6.3 In-context Learning with LLMs

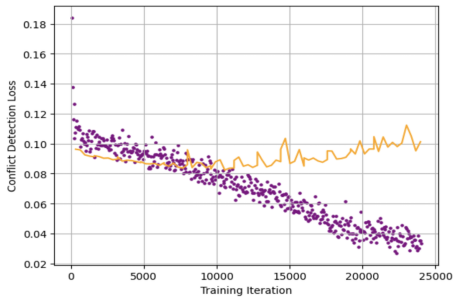
Following the idea of (Zhang et al., 2023), we generate demonstrations for story plausibility, conflict detection, action precondition prediction, and ac-



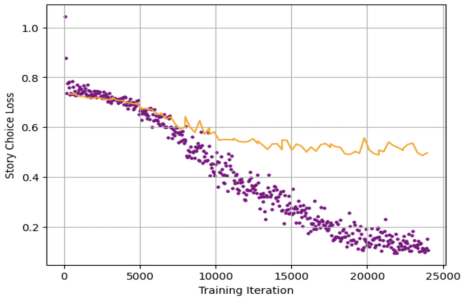
(a) Conflict detection loss in (Storks et al., 2021)



(b) Story choice loss in (Storks et al., 2021)



(c) Conflict detection loss under data augmentation



(d) Story choice loss under data augmentation

Figure 3: Training (purple, dotted) and validation (orange, solid) losses for best model trained on augmented TRIP dataset. Compared with (Storks et al., 2021).

tion effect prediction using four story pairs sampled from the training dataset. These demonstrations are integrated into the chain-of-thought prompting to guide GPT-3, Mistral-7B-Instruct-v0.3, and Llama-3.1-8B-Instruct to complete the four sub-tasks on the test dataset. Experimental results are presented in Table 5.

In-context learning with chain-of-thought prompting using large language models exhibits significant improvements. Specifically, GPT-3 increases consistency to 40.6% and verifiability to 10.1%, Mistral-7B increases consistency to 36.2% and verifiability to 11.5%, while Llama-3.1-8B achieves even better performance, reaching a consistency of 42.1% and a verifiability of 18.5%.

6.4 Meta-Learning

The meta-learning approach was evaluated for its ability to generalize across tasks with minimal fine-tuning. This method constructs support and query sets for each task, training the model to adapt quickly to unseen data.

The meta-learning method demonstrated strong performance in terms of accuracy and F1-score, as summarized in Table 6. With an average accuracy of 81.1% and an F1-score of 0.80, it outperformed baseline methods in scenarios requiring task generalization. This highlights the capability of meta-learning to leverage prior experience for rapid adaptation.

Two additional metrics, consistency and verifiability, were used to evaluate the logical coherence and reliability of predictions. The meta-learning approach achieved a consistency score of 28% and a verifiability score of 8%. Although these scores are lower compared to some baseline methods, they reflect the challenges inherent in generalization tasks where logical coherence and evidence-based predictions are more difficult to achieve.

The meta-learning approach excelled in scenarios involving limited labeled data and high variability across tasks. The results suggest that while the method effectively balances task-specific adaptation and generalization, it requires further refinement to improve its performance on metrics like consistency and verifiability. Future work could address this by integrating external knowledge bases or enhancing logical inference capabilities.

6.5 Comparison between Model Performance

Table 1 and Table 3 show the performance of various approaches on the test dataset, including advanced models, fine-tuning techniques, and diverse data augmentation strategies, all implemented with

Model	Accuracy (%)	Consistency (%)	Verifiability (%)
Baseline			
BERT	73.9	28.0	9.0
RoBERTa	73.6	22.4	10.6
ICL-CoT			
GPT-3	74.6	40.6	10.1
Mistral-7B	74.0	36.2	11.5
Llama-3.1-8B	74.4	42.1	18.5

Table 5: Experimental evaluation on the test dataset using in-context learning with LLMs

Model	Accuracy (%)	Consistency (%)	Verifiability (%)
Baseline			
BERT	73.9	28.0	9.0
RoBERTa	73.6	22.4	10.6
Meta-learning			
BERT	81.1	28.0	8.2

Table 6: Experimental evaluation on the test dataset using Meta-Learning

joint loss function. Fine-tuning pre-trained BERT on PIQA and RoBERTa on HellaSwag datasets shows modest accuracy improvements on the TRIP dataset (3% and 4% respectively over baseline), but results in decreased consistency and verifiability metrics. When trained on augmented datasets, the BERT model exhibits higher consistency and verifiability with a decrease in accuracy, while the RoBERTa model reaches higher accuracy and verifiability at the expense of consistency.

Table 2 and Table 4 present the performance of various approaches on the test dataset, including advanced models, fine-tuning techniques, and diverse data augmentation strategies, all implemented without story choice loss. Fine-tuning pre-trained BERT on PIQA and RoBERTa on HellaSwag datasets shows modest accuracy improvements on the TRIP dataset (2.63% and 3.89% respectively over baseline), but resulted in decreased consistency and verifiability metrics. For data augmentation techniques, BERT model achieves the highest verifiability of 11.7% when trained on the original dataset combined with the object-abstracted dataset, with little impact on the overall accuracy. Meanwhile, the RoBERTa model reaches the highest verifiability of 13.4% with the incorporation of information generated from knowledge graph.

Meta-learning, with a support-query framework, achieves strong accuracy (81.1%) while maintaining acceptable consistency (28%) and verifiability (8%). Its balance between accuracy and generalization highlights its potential for scenarios involving limited labeled data, although further refinement is needed to match the logical consistency and evidence-based reasoning achieved by some other methods.

Among all the approaches, in-context learning with chain-of-thought prompting using Llama-3.1-8B generates the best consistency of 42.1% and the best verifiability of 18.5%.

7 Discussion

7.1 Advanced Pre-trained Models and Fine-tuning

We conducted a comparative analysis of various model architectures, including a non-BERT architecture (Llama-2-7B) and two Transformer models (BERT and RoBERTa) pre-trained on specific datasets. The experimental results demonstrate that despite lacking task-specific pre-training, Llama-2-7B's robust architecture exhibits superior performance on the TRIP dataset, particularly in terms of verifiability and reasoning consistency. Specifically, BERT pre-trained on the PIQA dataset showed a 2.63% accuracy improvement, while RoBERTa pre-trained on the HellaSwag dataset achieved a 3.89% accuracy gain over the baseline. However, these pre-trained models, while improving accuracy, demonstrated significant degradation in model verifiability and reasoning consistency. This phenomenon may be attributed to the models' over-adaptation to the source datasets' feature distributions during pre-training, potentially

compromising their generalization capabilities on the TRIP dataset. Furthermore, Llama-2-7B maintained strong verifiability and consistency metrics while preserving high accuracy, suggesting that large language models may possess inherent advantages in handling reasoning tasks.

A noteworthy observation is that Llama-2-7B has clear architectural advantages over smaller Transformer-based models. Unlike BERT and RoBERTa, which rely heavily on task-specific pre-training, Llama-2-7B's scalability and extensive pre-training on different corpora appear to enable it to capture richer contextual and semantic representations. This is consistent with previous findings that large language models exhibit superior generalization capabilities in tasks involving multi-step reasoning and common-sense reasoning due to their increased capacity.

Furthermore, the results highlight the potential of leveraging pre-trained models such as Llama-2-7B for transfer learning without the need for task-specific pre-training. Their consistent performance across all metrics suggests that larger models with robust architectures can intrinsically overcome some of the limitations associated with overfitting specific feature distributions. This further raises the issue of diminishing returns of task-specific pre-training as model capacity increases, which requires deeper investigation in future work.

7.2 Data Augmentation

We have experimented with various data augmentation techniques including in-domain, external dataset, and knowledge graph information integration. It has been shown that increasing the size of the training dataset and incorporating additional information can lead to slight improvements in the verifiability of the models. Among all the experimented techniques, object abstraction appears to benefit BERT models the most, while integrating information generated from COMET helps improve the RoBERTa model performance the most. This suggests that simplifying entities into generalized categories and introducing potential action effects from knowledge graph may help models build better common-sense reasoning capabilities. However, the overall improvements remain modest, and the over-fitting issue remains unsolved. This indicates that data augmentation techniques are insufficient to tackle the fundamental issues of improving the models' reasoning abilities towards both low-level and high-level tasks.

In fact, there are some limitations that should be addressed in future work. For the synonym replacement based on WordNet, although we restricted the parts of speech and proportion of words that are replaced and performed some manual checks, it is inevitable that some synonyms may be inappropriate in the contexts, making the stories less plausible. Similarly, the knowledge graph information generated by COMET can sometimes be inaccurate or overly generic. These noises could potentially confuse the model, hindering its learning process. In addition, using COMET for knowledge graph information generation is time-consuming. Some more efficient and informative approaches need to be explored and experimented with. Moreover, the external dataset GITA is small in size and incorporating it into the training dataset seems to have little impact on model performance. Due to the limited time and resources, we have not yet made full use of larger corpora such as ROCStories, which require efforts to train models to automatically annotate preconditions and effects for each entity in each sentence. More efforts can be invested into this part in the future.

7.3 Prompting Large Language Models

In-context learning using large language models yields the best results among all approaches, but chain-of-thought prompting is a very traditional method. In fact, during the text generation process, a lot of free-form explanations for physical state prediction are meaningless. Since tasks like low-level physical state prediction cannot be further decomposed, the effect of guiding models to reason through low-level tasks may be constrained. This requires further investigation into more effective prompting techniques. For example, (Zhang et al., 2023) demonstrates that heuristic-analytic reasoning (HAR) for in-context learning with pre-trained language models can significantly improve the overall model performance on TRIP task. Future work could dive deeper into various prompting methods and integrating diverse reasoning structures to enhance model capabilities.

7.4 Meta-Learning

Another promising approach explored in this work is meta-learning, which employs a support-query framework to generalize across tasks. This method demonstrated an accuracy of 81.1% on the TRIP dataset, which is competitive with fine-tuning-based methods. However, its consistency (28%)

and verifiability (8%) lag behind those of in-context learning methods. The lower performance in these metrics indicates that while meta-learning effectively captures task adaptability, it currently struggles to align reasoning processes with real-world interpretability standards.

Future work could focus on refining meta-learning models by incorporating structured knowledge bases into the support-query framework or augmenting the approach with chain-of-thought prompting. These extensions may enhance both reasoning and interpretability, thereby improving the model's alignment with real-world requirements.

8 Conclusion

In this project, we explore ways to improve model performance on the TRIP dataset, a challenging benchmark for common-sense reasoning in physics. By applying advanced pre-trained models such as LLaMA-2-7B, fine-tuning using external datasets such as PIQA and HellaSwag, implementing data augmentation strategies, employing meta-learning frameworks, and prompting large language models for in-context learning, we aim to improve existing baselines in terms of accuracy, consistency, and verifiability.

Our experiments demonstrate that LLaMA-2-7B outperforms baseline models across all key metrics, achieving a 5.2% improvement in accuracy and a 7.8% gain in reasoning consistency compared to BERT. Data augmentation techniques, particularly object abstraction, are proved to be effective in improving BERT model performance, while knowledge integration from ConceptNet and COMET enhances RoBERTa models' ability to handle complex reasoning scenarios. However, synonym replacement and external datasets like GITA provide limited contributions due to their potential noises or size constraints. In-context learning with chain-of-thought prompting using Llama-3.1-8B generates the best result, achieving a consistency of 42.1% and a verifiability of 18.5%.

Despite the existing challenges, our results highlight the potential of in-context learning using large language models, meta-learning frameworks for task adaptability, as well as combining data augmentation, external knowledge, and advanced pre-trained models to enhance reasoning capabilities.

Future work could explore advanced augmentation strategies, such as dynamically generating augmented data using generative models or incorporating multimodal data to further enrich the training corpus. Refining the integration of knowledge graphs could involve leveraging more specialized resources, such as ATOMIC for causal reasoning or Wikidata for structured entity relationships. Additionally, adaptive weighting mechanisms could be applied to reduce the noise caused by less relevant knowledge triples.

Beyond improving performance on the TRIP dataset, the methods explored in this project could have broader applications in areas where physical common-sense reasoning plays a critical role, such as intelligent tutoring systems, conversational AI, and decision-support systems. By building on this foundation, we aim to contribute to the development of explainable and coherent AI systems capable of reasoning about physical common-sense.

9 Division of Work

Tianao Chen's Contribution

1. Reproduced RoBERTa, BERT, and DeBERTa models on the TRIP dataset using a joint loss function, along with a loss function that omits the story choice classification loss.

2. Trained the LLaMA-2-7B model on the TRIP dataset.

3. Fine-tuned the BERT model (trained on the PIQA dataset) and the RoBERTa model (trained on the HellaSwag dataset) on the TRIP dataset using a loss function that omits the story choice classification loss.

4. Wrote the related work, dataset, approaches, results, and discussion parts of the report.

5. Produced presentation slides.

Chuying Han's Contribution

1. Implemented all the data augmentation techniques on the TRIP dataset.

2. Trained BERT and RoBERTa models on all combinations of the augmented training datasets, both for all losses and omitting story choice loss.

3. Experimented with Mistral-7B for in-context learning using chain-of-thought prompting.

4. Contributed to the report by writing all parts related to data augmentation and in-context learning with LLMs in the sections of Related Work, Dataset, Approaches, Evaluation, Results, and Discussion.

5. Produced presentation slides.

Jiaqi Xu's Contribution

1. Trained and experimented Llama-3.1-8B using chain of thought prompting technique on TRIP dataset.

2. Trained and experimented with GPT3 using chain-of-thought prompting technique on the TRIP dataset.

3. Contributed to the report by writing sections related to the above experiments, abtract, part of introduction, part of related work.

4. Do the presentation.

Yukun Yang's Contribution

1. Implemented the meta-learning framework by designing and training tasks on support and query sets for the TRIP dataset.

2. Trained and evaluated the meta-learning approach with advanced pre-trained models.

3. Analyzed and compared the meta-learning results with other methods in terms of accuracy, consistency, and verifiability.

4. Write all things related to Meta-learning. Write or modify paragraphs in every section.

5. Do the presentation.

10 Code Repo

Our source code is publicly available at `https://github.com/cyhan192/CSE595_final_project_group4`.

References

Akari Asai and Hannaneh Hajishirzi. 2020. Logic-guided data augmentation and regularization for consistent question answering. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5642–5650, Online. Association for Computational Linguistics.

Emily M. Bender and Alexander Koller. 2020. Climbing towards NLU: On meaning, form, and understanding in the age of data. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5185–5198, Online. Association for Computational Linguistics.

Yonatan Bisk, Rowan Zellers, Ronan Le Bras, Jianfeng Gao, and Yejin Choi. 2019. Piqa: Reasoning about physical commonsense in natural language. *Preprint*, arXiv:1911.11641.

Antoine Bosselut, Hannah Rashkin, Maarten Sap, Chaitanya Malaviya, Asli Celikyilmaz, and Yejin Choi. 2019. Comet: Commonsense transformers for automatic knowledge graph construction. *arXiv preprint arXiv:1906.05317*.

Tom B Brown. 2020. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*.

Bhavana Dalvi, Lifu Huang, Niket Tandon, Wen-tau Yih, and Peter Clark. 2018. Tracking state changes in procedural text: a challenge dataset and models for process paragraph comprehension. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1595–1604, New Orleans, Louisiana. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. *Preprint*, arXiv:1810.04805.

Chelsea Finn, Pieter Abbeel, and Sergey Levine. 2017. Model-agnostic meta-learning for fast adaptation of deep networks. *Preprint*, arXiv:1703.03400.

Drew A. Hudson and Christopher D. Manning. 2019. Gqa: A new dataset for real-world visual reasoning and compositional question answering. *Preprint*, arXiv:1902.09506.

Filip Ilievski, Jay Pujara, and Hanzhi Zhang. 2021. Story generation with commonsense knowledge graphs and axioms. In *Workshop on Commonsense Reasoning and Knowledge Bases*.

Yifan Jiang, Filip Ilievski, and Kaixin Ma. 2023. Transferring procedural knowledge across commonsense tasks. In *ECAI 2023*, pages 1156–1163. IOS Press.

Justin Johnson, Bharath Hariharan, Laurens van der Maaten, Li Fei-Fei, C. Lawrence Zitnick, and Ross Girshick. 2016. Clevr: A diagnostic dataset for compositional language and elementary visual reasoning. *Preprint*, arXiv:1612.06890.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *Preprint*, arXiv:1907.11692.

Jieyi Long. 2023. Large language model guided tree-of-thought.

Junghyun Min, R Thomas McCoy, Dipanjan Das, Emily Pitler, and Tal Linzen. 2020. Syntactic data augmentation increases robustness to inference heuristics. *arXiv preprint arXiv:2004.11999*.

Roshanak Mirzaee, Hossein Rajaby Faghihi, Qiang Ning, and Parisa Kordjamshidi. 2021. SPARTQA: A textual question answering benchmark for spatial reasoning. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4582–4598, Online. Association for Computational Linguistics.

Bhavana Dalvi Mishra, Lifu Huang, Niket Tandon, Wen tau Yih, and Peter Clark. 2018. Tracking state changes in procedural text: A challenge dataset and models for process paragraph comprehension. *Preprint*, arXiv:1805.06975.

Nasrin Mostafazadeh, Nathanael Chambers, Xiaodong He, Devi Parikh, Dhruv Batra, Lucy Vanderwende, Pushmeet Kohli, and James Allen. 2016. A corpus and cloze evaluation for deeper understanding of commonsense stories. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 839–849.

OpenAI. 2023. Gpt-4 technical report. *ArXiv*, abs/2303.08774.

Giulia Pensa, Begoña Altuna, and Itziar Gonzalez-Dios. 2024. A multi-layered approach to physical commonsense understanding: Creation and evaluation of an Italian dataset. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 819–831, Torino, Italia. ELRA and ICCL.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2023. Exploring the limits of transfer learning with a unified text-to-text transformer. *Preprint*, arXiv:1910.10683.

Maarten Sap, Ronan Le Bras, Emily Allaway, Chandra Bhagavatula, Nicholas Lourie, Hannah Rashkin, Brendan Roof, Noah A. Smith, and Yejin Choi. 2019. Atomic: an atlas of machine commonsense for if-then reasoning. In *Proceedings of the Thirty-Third AAAI Conference on Artificial Intelligence and Thirty-First Innovative Applications of Artificial Intelligence Conference and Ninth AAAI Symposium on Educational Advances in Artificial Intelligence*, AAAI'19/IAAI'19/EAAI'19. AAAI Press.

Jake Snell, Kevin Swersky, and Richard S. Zemel. 2017. Prototypical networks for few-shot learning. *Preprint*, arXiv:1703.05175.

Robyn Speer, Joshua Chin, and Catherine Havasi. 2017. Conceptnet 5.5: an open multilingual graph of general knowledge. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence*, AAAI'17, page 4444–4451. AAAI Press.

Shane Storks, Qiaozi Gao, Yichi Zhang, and Joyce Chai. 2021. Tiered reasoning for intuitive physics: Toward verifiable commonsense language understanding. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 4902–4918, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Armin Toroghi, Willis Guo, Mohammad Mahdi Abdollah Pour, and Scott Sanner. 2024. Right for right reasons: Large language models for verifiable commonsense knowledge graph question answering. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing (EMNLP-24)*, Miami, USA. Association for Computational Linguistics.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2022. Self-consistency improves chain of thought reasoning in language models. *arXiv preprint arXiv:2203.11171*.

Jason Wei and Kai Zou. 2019. EDA: Easy data augmentation techniques for boosting performance on text classification tasks. 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 6382–6388, Hong Kong, China. Association for Computational Linguistics.

Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Thomas L. Griffiths, Yuan Cao, and Karthik R Narasimhan. 2023. Tree of thoughts: Deliberate problem solving with large language models. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. 2019. Hellaswag: Can a machine really finish your sentence? In *Annual Meeting of the Association for Computational Linguistics*.

Zhengyan Zhang, Xu Han, Zhiyuan Liu, Xin Jiang, Maosong Sun, and Qun Liu. 2019. ERNIE: Enhanced language representation with informative entities. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1441–1451, Florence, Italy. Association for Computational Linguistics.

Zheyuan Zhang, Shane Storks, Fengyuan Hu, Sungryull Sohn, Moontae Lee, Honglak Lee, and Joyce Chai. 2023. From heuristic to analytic: Cognitively motivated strategies for coherent physical commonsense reasoning. *arXiv preprint arXiv:2310.18364*.

Ben Zhou, Daniel Khashabi, Qiang Ning, and Dan Roth. 2019. "going on a vacation" takes longer than "going for a walk": A study of temporal commonsense understanding. *Preprint*, arXiv:1909.03065.