Three Approaches to Improve Reasoning on the TRIP Benchmark: Transfer Learning, Model Selection, and Prompting Techniques

Tiffany Parise, Yuting Duan, Xiyuan Chang, Junkuan Liu
University of Michigan
500 S State Street
Ann Arbor, MI 48109
{tparise, yutingd, xychang, junkuan}@umich.edu

Abstract

Given that Large Language Models (LLMs) are widely used in everyday society, it is important that these LLMs produce reliable, well-reasoned outputs. However, the TRIP benchmark reveals a concerning trend in which LLMs that purport high accuracy on reasoning tasks may not be able to justify their outputs with sound evidence. To address this issue, our project implements three approaches to improve reasoning abilities in LLMs and to encourage LLMs to generate their outputs by following coherent reasoning steps. Specifically, our three approaches include transferring knowledge from related reasoning tasks, employing powerful model architectures, and crafting prompts that surface reasoning abilities in LLMs. Through combinations of these approaches, we achieve approximately 20% improvements in performance on the lower-level reasoning tasks on the TRIP benchmark (Storks et al., 2021).

1 Introduction

As Large Language Models (LLMs) continue to rapidly grow in size and accuracy, they are being increasingly relied upon in our everyday society. Today, these LLMs are applied across a wide array of high-stakes industries, including medicine, law, and robotics (Kaddour et al., 2023). In fact, recent research estimates that LLMs will eventually impact a staggering 80% of workers in the United States (Eloundou et al., 2023). Outside of the workforce, the average person interacts with various LLMs and chatbots in their day-to-day life, with ChatGPT reaching an unprecedented 100M users within only two months of its release (Leng, 2024). The expanding reach of LLMs, across both our essential industries and our personal lives, makes it more important than ever for LLMs to produce reliable, well-reasoned outputs.

Unfortunately, prior work suggests that LLMs may lack these reasoning skills. For instance, LLMs are prone to hallucinations, where they invent inaccurate information that may appear to be true at face value (Huang et al., 2024). Even in cases where LLMs report high accuracy on reasoning tasks, some literature suggests that LLMs can artificially achieve this accuracy without actually performing the desired reasoning. Rather, the LLMs are exploiting spurious correlations that happen to exist in the training data (Chen et al., 2019; Gururangan et al., 2018), which means these LLMs may be unable to generalize to related reasoning tasks or to other plausible data distributions. This faulty reasoning and misleadingly high performance is concerning, especially given the extent to which our society relies on LLMs.

One key work that investigates this concern is the Tiered Reasoning for Intuitive Physics (TRIP) benchmark (Storks et al., 2021). In this benchmark, the authors ask LLM-based models to solve a standard commonsense reasoning task, with the caveat that the LLM must explicitly justify its output with evidence and reasoning. Specifically, the authors pose a story prediction task where the model is presented with two stories and must identify which story is plausible. To justify its choice, the model must examine the implausible story to determine which two sentences cause the implausibility. In even finer detail, the model must assess those sentences and their associated objects to identify which physical states cause the sentences to conflict. Figure 1 presents an example of two input stories and the tiered process the model follows to evaluate those stories.

Through this tiered three-step evaluation, the TRIP benchmark allows us to measure not only the model's accuracy in a high-level reasoning task (story prediction), but also the extent to which this high-level output is grounded in logical reasoning about relevant lower-level subtasks (conflict detection and physical state detection). Unfortunately, when this evaluation is performed on BERT-based

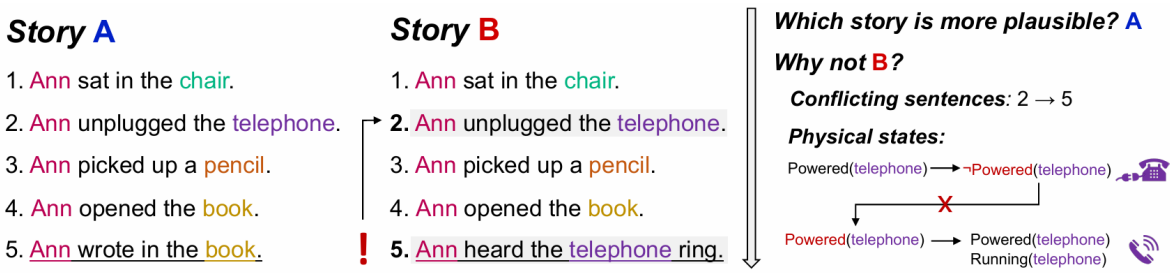

Figure 1: An example input and output in the TRIP benchmark. Given two stories, the model must identify which story is plausible, identify the two conflicting sentences in the implausible story, and identify the physical state changes in the conflicting sentences. This tiered evaluation assesses the model's reasoning skills on both a high-level task (story prediction) and its lower-level subtasks (conflict detection and physical state detection). Image from (Storks et al., 2021)

models (Devlin et al., 2019; Liu et al., 2019; He et al., 2020), it reveals a concerning trend where the performance on the high-level task significantly exceeds the performance on the two lower-level subtasks, as shown in Table 1. This trend indicates that the model's true reasoning ability is actually much lower than the high accuracy may have led us to believe.

Given the concerning gap between the models' performance on the high-level and low-level reasoning tasks, our project implements three approaches to improve the models' lower-level reasoning skills and to encourage the model to apply these skills when solving the high-level task. These approaches include transfer learning, model selection, and advanced prompting.

In our first approach, we aimed to improve the models' lower-level reasoning skills via transfer learning. We identified relevant commonsense reasoning datasets that emphasize physical state detection and conflict detection, and we fine-tuned the model on these datasets to improve its reasoning skills in these low-level areas. Our performance on these fine-tuned models was comparable or slightly reduced from the original TRIP benchmark. This result may support concerns that models trained on commonsense reasoning datasets have not learned generalizable reasoning skills, but rather learned spurious correlations for solving that specific end task.

Given that the BERT-based model did not display sufficient low-level reasoning even after fine-tuning, our second approach replaced the BERT-based model with a more powerful LLM that contains significantly more parameters and may therefore be better equipped for complicated reasoning tasks. Our evaluation tests a variety of models that differ by developer (LLaMA vs. Mistral), number of parameters (7, 8, or 13 billion), instruction fine-tuning (included or not), and version (LLaMA 2 vs. LLaMA 3) in order to isolate which aspects of these models improve performance the most. Our best-performing model is Mistral-7B-Instruct-v0.3, which achieves strong performance of 40.14% consistency and 27.46% verifiability on the two low-level reasoning tasks. This model also had the lowest variance and the highest accuracy (61.97%) on the high-level task, which suggests that strong low-level reasoning is a more successful and reliable path to high end-task performance.

Our third approach aims to further improve the performance of Mistral-7B-Instruct-v0.3 by applying targeted prompts that encourage the model to reason. Especially because many of our evaluated models observed high variance across repeated trials, we hypothesize that these models do have the necessary reasoning skills to perform well on the TRIP benchmark, but they struggle to access these skills in practice. Thus, our third approach assesses the extent to which few-shot demonstrations and role-playing prompts can surface the model's commonsense reasoning. Our results show that model performance can vary significantly as we change the number of demonstrations and their represented conflict types and object types. By carefully assessing these variations, we curated sets of few-shot demonstrations that improve accuracy and consistency by almost 10%. Further, we were able to design a role-play prompt that successfully enhanced the model's physical state understanding and improved its performance on the verifiability metric.

Overall, the key contribution of our work is a comprehensive evaluation of how transfer learning, model selection, and prompting techniques can be

Model	Accuracy ↑	Consistency ↑	Verifiability ↑
Random	49.5%	10.7%	0.0%
BERT	70.9%	21.9%	8.3%
ROBERTA	**72.9%**	19.1%	**9.1%**
DEBERTA	**72.9%**	**22.2%**	6.6%

Table 1: Performance of BERT-based models on the TRIP benchmark (Storks et al., 2021). Accuracy measures performance on story prediction. Consistency measures performance on both story prediction and conflict detection. Verifiability measures performance on story prediction, conflict detection, and physical state detection.

used to improve a model's lower-level reasoning skills and its ability to apply these skills in practice. Our most successful configuration is a Mistral-7B-Instruct-v0.3 model with few-shot prompting, which improves the TRIP benchmark's consistency and verifiability by approximately 20% compared to the original (Storks et al., 2021) paper.

2 Related Work

This section contextualizes our project against related work in commonsense reasoning. First, we discuss standard commonsense benchmarks and how they differ from the tiered evaluation in the TRIP benchmark. Next, we discuss prior works that transfer knowledge across commonsense reasoning tasks, which motivates our project's transfer learning approach. Then, we discuss SOTA models that are tailored for reasoning tasks and how these models compare to the LLMs our project evaluates. Finally, we discuss prompting strategies our project applies to improve performance on the TRIP benchmark.

2.1 Standard Datasets in Commonsense Reasoning

There are several popular benchmarks that are used to train and evaluate commonsense reasoning in LLMs, where commonsense reasoning generally refers to a model's ability to apply widely-accepted logic in the context of common events and occurrences. These datasets encompass a wide range of commonsense areas, including physical (Bisk et al., 2019), cultural (Shen et al., 2024), temporal (Mostafazadeh et al., 2016), and spatial (Mirzaee et al., 2021), among others. Our project specifically focuses on physical commonsense reasoning, which is traditionally challenging for LLMs because they exist online and primarily observe our 3D physical world secondhand through text.

2.2 Tiered Datasets for Procedural Reasoning

One challenge with commonsense reasoning benchmarks is that many of them only evaluate the model's accuracy on a specific end task. In these cases, the model may achieve artificially high accuracy by exploiting spurious correlations that exist in the data, as opposed to performing actual human-level reasoning (Gururangan et al., 2018). Some datasets (Chen et al., 2019) address this by incorporating adversarially designed examples that models cannot answer correctly if they overfit to spurious correlations. Other datasets (Storks et al., 2021; Tandon et al., 2018; Yang et al., 2018) incorporate a tiered evaluation process where the model must justify its end-task output by correctly solving lower-level subtasks of the end-task. The intent of such datasets is to identify unwanted performance gaps between the model's high-level and low-level reasoning. Our project uses the tiered evaluation framework presented in the Tiered Reasoning for Intuitive Physics (TRIP) benchmark (Storks et al., 2021).

2.3 Transferring Knowledge from Other Datasets

Given the wide range of available commonsense reasoning datasets, researchers have turned to transfer learning as a strategy to increase model generalizability and improve end-task performance. For example, (Jiang et al., 2023b) transfers knowledge from a complicated procedural reasoning dataset to learn sound reasoning processes and therefore improve performance on a variety of simpler commonsense reasoning benchmarks. Models may also transfer information from knowledge graphs (Sap et al., 2019; Speer et al., 2018) that encode more broadly applicable reasoning trends compared to the task-specific patterns in many commonsense reasoning datasets. There are even works that can automatically construct these knowledge bases (Bosselut et al., 2019). In our project, we apply

transfer learning to improve performance on the TRIP benchmark (Storks et al., 2021). Specifically, we transfer knowledge from commonsense reasoning datasets that specifically target the low-level reasoning skills the TRIP benchmark underperforms on.

2.4 Tailored Architectures for Reasoning

In addition to *datasets* that emphasize procedural reasoning structures, there are also *models* and associated learning frameworks that do this. For example, the CGLI model (Ma et al., 2022) emphasizes consistency between local and global patterns, and the Breakpoint Transformer (Richardson et al., 2022) emphasizes reliable reasoning throughout intermediate steps in the model. Both of these models achieve near SOTA performance on the TRIP benchmark. In comparison to these models that are specifically designed to reason well, our project evaluates the zero-shot reasoning capabilities of general-purpose LLMs.

2.5 Pretrained Language Models

Pretrained Language Models (PLMs) have shown impressive performance on commonsense reasoning tasks. In particular, models in the LLaMA (Dubey and et al., 2024), Mistral (Jiang et al., 2023a), and GPT (Ouyang et al., 2022) families have reported significant improvements in accuracy for a variety of commonsense reasoning tasks. Given the significant advancements in LLMs over the past few years, the BERT-based PLMs evaluated in the 2021 TRIP benchmark (Storks et al., 2021) are now outdated. Therefore, our project replaces these BERT-based PLMs with more recent models. To identify exactly which components of these models improve performance the most we evaluate 6 PLMs that vary in their developer, version, parameter count, and instruction fine-tuning (Zhang et al., 2024). Further, our approach prioritizes the evaluation of small language models (ranging from 7B to 13B parameters) in response to growing concerns about large language models that have high carbon footprints, low interpretability, and unequal accessibility across researchers (Bender et al., 2021).

2.6 Variance in LLM Outputs

Although LLMs purport high accuracy across several commonsense reasoning tasks, one issue is that this accuracy may not be consistent across repeated trials on the same prompts and datasets. In fact, recent work (Atil et al., 2024) suggests that LLM accuracy can vary as much as 10%, even when the LLM's configuration is not changed across runs. In response to this concern, our project analyzes the margin of error of various LLMs on the TRIP benchmark. Among our evaluated models, the Mistral models have the most consistent performance. Additionally, the model with the lowest variance also achieves the highest performance on the two low-level reasoning tasks in the benchmark, which indicates that models following structured, logical reasoning processes can more consistently arrive at the correct output.

2.7 Prompts that Encourage Reasoning

In addition to the accuracy variations mentioned above, LLM performance is also sensitive to the specific phrasing of the input prompt. In fact, many studies have identified methods of writing input prompts that encourage model reasoning and thus significantly improve performance. For example, chain-of-thought prompts (Wei et al., 2022) explicitly ask the model to explain its reasoning, which helps the model identify and correct its own mistakes. As another strategy, few-shot prompting (Brown, 2020) provides the model with examples of the desired output format, which allows the model to quickly adapt to new tasks. In addition, role-play prompting (Kong et al., 2023) improves performance by asking the model to respond in the context of a specific persona who may be more likely to follow the desired reasoning processes. Finally, self-consistency prompting (Wang et al., 2023) allows the model to generate more reliable outputs by considering and combining multiple potential responses to the prompt. Our project applies many of these prompting strategies in an effort to encourage LLMs to perform deeper reasoning on the TRIP benchmark.

3 Description of Approaches

In this section, we describe the three approaches we implemented to improve performance on the TRIP benchmark: transfer learning, model selection, and advanced prompting techniques. In particular, our goal is to improve the lower-level reasoning skills of the evaluated LLMs and to improve these LLMs' ability to utilize those skills in practice. For each approach, we describe our motivation, design choices, and implementation details.

3.1 Approach 1: Transfer Learning with BERT

Our first approach to improve performance on the TRIP benchmark is to apply transfer learning to the BERT-based model used in the original TRIP paper (Storks et al., 2021).

3.1.1 Review of BERT in TRIP

BERT (Bidirectional Encoder Representations from Transformers) (Devlin et al., 2019) is a state-of-the-art transformer-based language model developed by Google. It is designed to understand the context of words in a text by considering the words that come before and after them (bidirectional context), which makes it powerful for various natural language processing (NLP) tasks. Unlike traditional models that process text in a unidirectional manner (e.g., left-to-right or right-to-left), BERT processes text bidirectionally. This means it learns contextual word representations by looking at the entire sequence of words, both preceding and following a target word, during training. BERT is based on the Transformer model, specifically its encoder mechanism. At a high level, BERT consists of four modules: tokenizer, embedding, encoder and task head.

BERT was pre-trained on two tasks, which are masked language modeling and next sentence prediction. BERT can be fine-tuned with fewer resources on smaller datasets to optimize its performance on specific tasks such as natural language inference and text classification, and sequence-to-sequence-based language generation tasks such as question answering and conversational response generation.

The TRIP paper (Storks et al., 2021) utilizes the large version of BERT (BERT-Large), which includes 24 layers, 16 attention heads and 1024 hidden units with 355 M parameters in total. The benchmark with BERT in the TRIP paper has good accuracy performance but concerning results in verifiability and consistency, as Table 1 shows.

3.1.2 Review of Transfer Learning

Transfer learning (Bozinovski, 2020) is a transformative approach in machine learning, particularly impactful in natural language processing (NLP), where pre-trained models like BERT (Devlin et al., 2019), GPT (Ouyang et al., 2022), and LLaMA (Dubey and et al., 2024) have revolutionized tasks such as text classification, translation, and question answering. By leveraging knowledge gained from large-scale corpora, these models require minimal labeled data to adapt to target tasks, significantly reducing training time and improving performance. Common techniques include fine-tuning pre-trained models and adapting them through domain-specific training. Despite its success, challenges like domain mismatches, overfitting on small datasets, and inherent biases in pre-trained models persist as (Zoph et al., 2020) points out.

3.1.3 Datasets for Transfer Learning

We transfer knowledge from two datasets (Conversational Entailment and Physical Interaction Question Answering) that emphasize the low-level reasoning skills that the TRIP benchmark currently underperforms on.

Conversational Entailment (CE (Zhang and Chai, 2010)) was first introduced by Chen Zhang and Joyce Chai in 2010. It is designed to evaluate a model's ability to perform entailment reasoning within the context of dialogues or conversations. Conversational entailment is a subtask of natural language inference that involves determining whether a hypothesis logically follows (entailment), contradicts (contradiction), or is neutral to a premise, where the premise typically arises from conversational exchanges. These reasoning skills align well with the story plausibility and conflict detection tasks in the TRIP benchmark.

Physical Interaction Question Answering (PIQA (Bisk et al., 2019)) dataset is a benchmark designed to evaluate a model's ability to reason about physical interactions in everyday scenarios. It focuses on common-sense physical reasoning, which involves understanding how objects and actions work together in the real world. As the task of PIQA, given a natural language question and two possible answers, the model must choose the more plausible answer based on physical reasoning. By fine-tuning on PIQA, we aim to improve the TRIP benchmark's performance on physical state detection.

The motivation is to enable the BERT model to learn reasoning skills through transfer learning from the CE and PIQA datasets. Although the original TRIP paper (Storks et al., 2021) evaluates BERT, ROBERTA, and DEBERTA, our evaluation focuses on BERT because it had the most balanced performance on the consistency and verifiability metrics.

	Accuracy ↑	Consistency ↑	Verifiability ↑
BERT	70.9%	21.9%	8.3%
BERT Fine-tuned on CE	**72.9%**	21.9%	5.4%
BERT Fine-tuned on PIQA	70.4%	16.8%	5.4%

Table 2: Performance on TRIP Benchmark After Transfer Learning

3.1.4 Detailed Transfer Learning Implementation

The large version of BERT was fine-tuned on the PIQA and CE datasets to enhance its performance on the TRIP dataset. The finetuned BERT models with PIQA and CE model were saved locally, with embeddings, configuration files, pretrained model weights, and the tokenizer, to ensure efficient and customized adaptation. The finetuned model with local files was then trained and tested on TRIP with the three metrics, accuracy, consistency, and verifiability. By leveraging the pre-trained knowledge from BERT and refining it with domain-specific datasets, this approach aims to improve the model's ability to generalize across related tasks while maintaining robust contextual understanding.

3.2 Approach 2: Use More Powerful Large Language Models

Given that the BERT-based model in Approach 1 did not display sufficient low-level reasoning even after fine-tuning, our next approach replaces this BERT-based model with compact state-of-the-art models, Mistral and LLaMA, to investigate their potential in addressing the multi-tiered reasoning challenges posed by the TRIP dataset. In this section, we detail our methodological advancements in applying in-context learning (ICL) strategies to these LLMs for commonsense reasoning tasks, particularly emphasizing the integration of heuristic-analytic reasoning (HAR).

3.2.1 Heuristic-Analytic Reasoning in Contextual Language Models

Now that we have switched from a BERT-based model to a LLM that accepts free-text inputs, we need to identify a strategy for phrasing our prompts to the LLMs. The state-of-the-art prompts for the TRIP benchmark follow the heuristic-analytic reasoning process outlined in (Zhang et al., 2023).

The concept of heuristic-analytic reasoning (HAR) is rooted in dual-process theories from cognitive psychology, where humans combine fast, intuitive heuristic thinking with slower, deliberative analytic reasoning to make decisions and rationalize them coherently. Inspired by this cognitive framework, (Zhang et al., 2023) designed a methodology to guide pre-trained language models (PLMs) through multi-tiered reasoning tasks. HAR provides a structured mechanism by which higher-level heuristic decisions inform and refine lower-level analytic rationalizations.

Within the context of the TRIP dataset, HAR evaluates reasoning coherence through tasks requiring both high-level decisions, such as story plausibility classification, and granular evidence-based justifications, such as identifying conflicting sentences and underlying physical states. HAR addresses these requirements by prioritizing heuristic processes for initial decision-making, followed by analytic processes to ensure detailed and verifiable reasoning.

In addition to applying HAR, the prompts in (Zhang et al., 2023) also leverage in-context learning. In-context learning allows PLMs to adapt to specific tasks without gradient updates by leveraging demonstration examples provided in the input prompt. This approach guided the PLM through top-down reasoning, beginning with high-level predictions and iteratively refining context relevance at each reasoning step.

3.2.2 Experimental Setup and Model Selection

Existing research in commonsense reasoning often emphasizes larger models, such as GPT-3 or LLaMA-65B, due to their high accuracy in downstream tasks. However, these larger architectures are computationally intensive, limiting their practical application. By contrast, compact models present an opportunity to achieve similar coherence metrics with reduced resource requirements. Furthermore, the limited exploration of small models within the domain of reasoning tasks creates a compelling gap that our research addresses. Therefore, in implementing HAR, we focused on the compact PLMs listed in Table 3.

Our experimental pipeline utilized in-context learn-

Model	# Parameters	Accuracy ↑ (%)	Consistency ↑ (%)	Verifiability ↑ (%)
BERT (Storks et al., 2021)	355M	70.9	21.9	8.3
ROBERTA (Storks et al., 2021)	355M	72.9	19.1	9.1
DEBERTA (Storks et al., 2021)	140M	72.9	22.2	6.6
LLaMA-2	7B	51.83 ± 2.07	17.11 ± 1.90	5.77 ± 0.91
LLaMA-2	13B	50.85 ± 1.82	21.06 ± 1.83	11.62 ± 1.39
LLaMA-3.1	8B	52.04 ± 3.86	31.13 ± 3.46	23.80 ± 2.80
LLaMA-3.1-Instruct	8B	52.75 ± 2.53	31.90 ± 3.45	20.14 ± 1.99
Mistral-v0.3	7B	53.59 ± 0.16	31.76 ± 0.16	23.31 ± 0.16
Mistral-Instruct-v0.3	7B	61.97 ± 0.00	40.14 ± 0.00	27.46 ± 0.00
LLaMA (Zhang et al.)	65B	55.6	44.4	35.2
InstructGPT (Zhang et al.)	175B	72.6	47.9	23.9

Table 3: Performance comparison of models on the TRIP dataset using accuracy, consistency, and verifiability metrics. Higher values indicate better performance. Our results are reported with a 95% confidence interval across 10 trials for each model.

ing, wherein demonstration examples were concatenated to input prompts to adapt the models to TRIP tasks without gradient updates. This design allowed us to assess the efficacy of HAR in improving the reasoning capabilities of compact models, particularly their ability to align heuristic and analytic reasoning steps.

3.3 Approach 3: Incorporating Advanced Prompting Techniques

The TRIP benchmark performance in (Storks et al., 2021) indicates a concerning lack of reasoning on the lower-level tasks. Thus, we incorporated advanced prompting techniques to enhance the reasoning capabilities of our evaluated PLMs and maximize the effectiveness of in-context learning (ICL). These techniques aim to refine model behavior by manipulating the structure, content, and diversity of demonstrations provided during inference. This section details our exploration of three key strategies: optimizing the number of demonstrations, analyzing the variance introduced by demonstration selection, and employing role-playing prompts to guide reasoning.

3.3.1 Number of Demonstrations

The number of demonstrations included in a prompt plays a crucial role in determining the quality of the model's predictions. Intuitively, a higher number of demonstrations should provide more context for the model to generalize its reasoning processes. However, this is counterbalanced by practical constraints such as the fixed context length of PLMs, especially in smaller architectures like Mistral-7B and LLaMA-3.1-8B-Instruct.
In our experiments, we systematically varied the number of demonstrations to evaluate their impact on task performance across TRIP's multi-tiered reasoning metrics. We observed diminishing returns beyond a certain number of demonstrations, likely due to the model's difficulty in maintaining focus across longer input sequences. These findings motivate us to build a high quality prompt by strategically choosing demonstrations.

3.3.2 Demonstration Selection

The selection of demonstrations in few-shot learning settings significantly impacts the model's performance, particularly in tasks requiring nuanced reasoning. To investigate the effects of demonstration selection on the TRIP benchmark, we conducted experiments that varied two key aspects of the demonstrations: the type of conflict in the implausible story, and the objects and states in the stories.

Demonstrations with Different Conflict Types
Conflict types within the TRIP dataset can be broadly categorized into explicit and implicit. Explicit conflicts arise from direct contradictions in the narrative, such as an object being presented as "unedible" in one sentence but "edible" in the next.

In contrast, implicit conflicts rely on unstated assumptions or commonsense reasoning, such as an object being described as cold despite conditions that indirectly imply this is not the case. Figure 2 provides a more detailed example of these two conflict types.

To understand the impact of varying conflict types on model performance, we tested five settings for the distribution of explicit and implicit conflicts in the few-shot demonstrations. The goal was to evaluate how these distributions influence the model's ability to accurately predict story plausibility and maintain coherence in its reasoning.

Demonstrations with Different Object Types We also measure the variance caused by changing which objects are included in the few-shot examples. The goal of this experiment is to evaluate how robust the model is to the specific objects described in the input stories.

Specifically, given an initial set of few-shot demonstrations, we generate a new set of demonstrations by modifying the object that causes conflict to another similar object. We then evaluate the model on both this modified set and the original set to compare their performance. For example, given a demonstration containing the sentence "Mary ate the donut", our replacement sentence could be "Mary ate the banana." Because both the original and replaced objects are foods, these sentences are conceptually equivalent, yet we discover that this substitution does in fact impact model performance.

Through this approach, we aim to understand how variations in object types and demonstrations affect accuracy, consistency, and verifiability, as well as to identify the specific configurations that lead to improved performance on the TRIP benchmark. This prompting strategy provides insights into the role of few-shot examples in enhancing reasoning tasks.

3.3.3 Role-Playing Prompts

In reasoning tasks that can be approached from multiple angles, role-playing prompts guide model behavior by assigning the model to predefined role that will tend to approach the task in a particular way. Our project designed and evaluated two roles: "Careful Story Editor" and "Interior Decorator." The careful story editor is described as meticulously analyzing narratives for logical consistency, while the interior decorator is described as intuitively understanding physical objects and their common usages. In this way, we expect that the careful story editor role will improve consistency, while the interior decorator role will improve verifiability.

4 Evaluation

In this section, we discuss our quantitative results for each of the three approaches described in the previous section. In our evaluation, we emphasize comprehensive and robust analysis. Based on this analysis, we make continual improvements to our project strategy across our three approaches.

4.1 Metrics for Multi-Tiered Reasoning

To evaluate the reasoning coherence of our models, we utilized the three-tiered evaluation metrics defined by the TRIP benchmark: accuracy, consistency, and verifiability. Each metric captures a distinct aspect of reasoning:

- **Accuracy** measures the correctness of high-level decisions, specifically identifying the plausible story from a pair of narratives.
- **Consistency** evaluates whether the model correctly identifies the conflicting sentences within the implausible story, linking high-level decisions to specific textual evidence.
- **Verifiability** assesses the model's ability to justify conflicts through correct predictions of underlying physical states. This metric ensures that the reasoning chain is coherent and fully supported by evidence.

These metrics collectively provide a comprehensive framework for assessing not just end-task performance but also the logical rigor of the model's reasoning processes.

4.2 Transfer Learning Results

After hyperparameter tuning, we pre-trained the BERT model on the CE dataset using a batch size of 1, learning rate of 10^{-5} and 10 epochs, and we pre-trained on the PIQA dataset using a batch size of 8, learning rate of 10^{-5} and 10 epochs.

Table 2 compares the performance of the original large BERT model to the performances of the BERT models we fine-tuned on the CE and PIQA

Explicit conflicts	
Mary tossed the donut in the trash	effect: donut is inedible
Mary ate the donut	precondition : donut is edible

Implicit conflicts	
Tom put the soup in the microwave	implies the soup should be heated up
Tom ate the cold soup	precondition : soup is cold

Figure 2: Examples of the two types of conflicts in the implausible stories in the TRIP dataset

datasets. The finetuned model from CE can increase accuracy. However, the benchmarks of the low-level tasks, consistency and verifiability, do not have better performance after transfer learning with the two datasets, CE and PIQA.

4.3 Model Selection Results

Table 3 provides a comprehensive comparison of six pre-trained language models (PLMs) evaluated on the TRIP dataset. For each model, we report its performance with 95% confidence, as measured across 10 trials per model. These results reveal several interesting patterns regarding model performance and their alignment with multi-tiered reasoning requirements.

Performance of Mistral vs. LLaMA Models A notable observation from the results is the superior performance of the Mistral series over the LLaMA series across most metrics. For example, Mistral-7B-Instruct-v0.3 achieves the highest scores among smaller models for both consistency (40.14%) and verifiability (27.46%). In contrast, LLaMA-3.1-Instruct lags behind, with consistency at 31.90% and verifiability at 20.14%. Similarly, even the non-instruction-finetuned version, Mistral-7B-v0.3, outperforms both LLaMA-3.1-8B and LLaMA-3.1-8B-Instruct in these two metrics.
This trend suggests that the Mistral models are inherently better at aligning high-level predictions with specific evidence and providing justifications, likely due to their architecture or fine-tuning strategies. The instruction-tuned variant of Mistral further amplifies these advantages, particularly in reasoning coherence as measured by consistency and verifiability.

Relationship Between Performance and Variance Another critical insight from the results is the relationship between a model's performance on consistency and verifiability and its variance across trials. Models that perform better on these metrics, such as Mistral-7B-Instruct-v0.3, exhibit lower variance. For instance, Mistral-7B-Instruct-v0.3 demonstrates near-zero variance for both consistency (40.14 ± 0.00%) and verifiability (27.46 ± 0.00%), indicating its robustness across multiple trials. In contrast, LLaMA-3.1-8B-Instruct, which performs worse on these metrics, exhibits higher variance (31.90 ± 3.45% for consistency and 20.14 ± 1.99% for verifiability). This observation suggests that models capable of maintaining reasoning coherence are less sensitive to the specific configurations of the task or the inherent randomness in few-shot learning setups.

4.4 Prompting Strategy

We explore the variance of LLM outputs with different prompting strategies. In detail, we investigate the impacts of number of demonstrations, demonstration selection with different conflict types and object types and role-playing prompting on Mistral model.

4.4.1 Number of Demonstrations

Two models Mistral-7B-Instruct-v0.3 and Mistral-7B-v0.3 with 2, 4 and 6 few-shot demonstrations are evaluated. The 4 demonstrations correspond to the default examples provided in the code (Zhang et al., 2023), while the 2 demonstrations are the first two examples from this set. The 6 demonstrations represent the examples selected for the default balanced option. As the results shown in Table 6 and Table 7, both models achieved the best overall performance with 4 demonstrations, while their performance were worst with 6 demonstrations. This suggests that using 6 demonstrations may lead to overfitting, resulting in poorer performance.

4.4.2 Demonstration Selection

Demonstrations with Different Conflict Types The results of the two models, Mistral-7B-Instruct-v0.3 and Mistral-7B-v0.3, as shown in Tables 8

Prompting Strategy	Accuracy ↑	Consistency ↑	Verifiability ↑
HAR	61.97%	40.14%	27.46%
HAR with Careful Story Editor Role	61.27%	40.14%	26.76%
HAR with Interior Decorator Role	62.68%	38.03%	28.17%

Table 4: Performance of Mistral-7B-Instruct-v0.3 with role-playing prompts

Prompting Strategy	Accuracy ↑	Consistency ↑	Verifiability ↑
HAR	53.59%	31.76%	23.31%
HAR with Careful Story Editor Role	55.63%	34.51%	27.46%
HAR with Interior Decorator Role	55.63%	35.21%	27.46%

Table 5: Performance of Mistral-7B-v0.3 with role-playing prompts

and 9, reveal several notable trends regarding the relationship between conflict type distributions in demonstrations and model performance across accuracy, consistency, and verifiability.

- *Best Results Do Not Occur in All-Explicit or All-Implicit Configurations*
  The best performance in terms of consistency and verifiability does not occur at the extremes of conflict distributions, such as 4 explicit conflicts and 0 implicit conflicts, or 0 explicit conflicts and 4 implicit conflicts. Instead, configurations with a balanced mix of explicit and implicit conflicts tend to yield better results. This suggests that a balance between explicit and implicit conflicts provides the model with richer yet interpretable demonstrations, enabling better alignment of high-level predictions with underlying evidence. All-explicit configurations may oversimplify the task, providing limited opportunities for the model to generalize beyond simple, surface-level contradictions. On the other hand, all-implicit configurations often demand nuanced commonsense reasoning and rely on implicit assumptions that the model may not be well-equipped to handle, leading to decreased performance in consistency and verifiability.

- *Instruction-Tuned Models Perform Better in the Consistency Metric*
  Across all conflict type distributions, Mistral-7B-Instruct-v0.3 consistently outperforms Mistral-7B-v0.3 in consistency. This improved performance in consistency can be attributed to instruction fine-tuning, which enhances the model's understanding at the sentence level (Zhang et al., 2024). Instruction tuning encourages the model to align its predictions more closely with sentence-level evidence, enabling it to better identify explicit conflicts and maintain coherence when reasoning across multiple sentences.
  However, since instruction fine-tuning is trained at the sentence-level, it does not appear to significantly improve the model's understanding at the physical state level, which is a critical factor for verifiability. Verifiability requires a deeper reasoning process that connects high-level predictions with detailed physical state transitions, such as preconditions and effects. The lack of improvement in verifiability scores, and the fact that Mistral-7B-Instruct-v0.3 does not consistently outperform Mistral-7B-v0.3 on this metric, supports that instruction tuning primarily benefits tasks requiring sentence-level reasoning rather than tasks requiring deeper physical commonsense understanding.

Demonstrations with Different Object Types
In our experiments, we observe significant variance in performance metrics when replacing the objects in the few-shot demonstrations with conceptually similar alternatives. As shown in Tables 10 and 11, we first select a set of default few-shot demonstrations (Initial Demo #1) and generate a modified set (Modified Demo #1) by replacing objects that cause conflicts with similar objects. Additionally, we select a second set of few-shot demonstrations (Initial Demo #2) and evaluate the model's performance on both this set and its modified version (Modified Demo #2). This setup allows us to measure how object replacements affect accuracy, consistency,

Number of Demonstrations	Accuracy ↑	Consistency ↑	Verifiability ↑
2	62.67%	37.32%	20.42%
4	61.97%	40.14%	27.46%
6	54.23%	37.32%	15.49%

Table 6: Performance of Mistral-7B-Instruct-v0.3 with Different Numbers of Few-Shot Demonstrations

Number of Demonstrations	Accuracy ↑	Consistency ↑	Verifiability ↑
2	49.30%	30.99%	21.13%
4	53.59%	31.76%	23.31%
6	50.00%	35.21%	10.56%

Table 7: Performance of Mistral-7B-v0.3 with Different Numbers of Few-Shot Demonstrations

and verifiability for the Mistral-7B-Instruct-v0.3 and Mistral-7B-v0.3 models.

In this analysis, we examined the impact of modifying the object types in few-shot demonstrations on the model's performance. Our initial hypothesis was that the reasoning process should remain object-agnostic, meaning that replacing the objects causing conflicts with different types of objects would not affect the model's performance. However, the results in Table 10 and 11 reveal unexpected changes in accuracy, consistency, and verifiability after object replacement. For example, for Mistral-7B-v0.3-Instruct, accuracy improves slightly (+0.99%) in the first set but decreases significantly (-4.22%) in the second set. This variability indicates that while object replacement may provide richer context for reasoning, it can also introduce additional complexity that challenges the model's ability to make accurate predictions.

Intuitively, the object substitutions should not change the performance because the logic and plausibility of the input stories has not changed. Specifically, all of our object replacements are conceptually equivalent, such as replacing "donut" with another food like "banana" or "apple". However, we still observe performance changes of as much as 7%, which may suggest that the models are sensitive to subtle contextual shifts in the narrative. Alternatively, it is possible that the objects in the TRIP dataset do not all occur at equal rates, such that more frequent objects may be more useful as examples.

In summary, the demo selection plays an essential role in model performance. Additionally, the experiments reveal that object replacements introduce variance in performance metrics due to subtle narrative shifts, even for conceptually equivalent objects. This highlights the importance of carefully selecting few-shot demonstrations to ensure optimal model consistency with reasoning tasks like TRIP.

4.4.3 Role-Playing Prompts

Role-Playing prompts generally improved performance on both of the evaluated Mistral models, as shown in Tables 4 and 5. The performance improvements are particularly noticeable for the Mistral-7B-v0.3 model, whose performance increased by 2-4% for each metric. Notably, the Interior Decorator role improved accuracy and verifiability on both models. This makes sense because the Interior Decorator was described as having a strong understanding of physical states, which is directly aligned with the verifiability metric. We also observed that the Careful Story Editor did not improve performance to the extent of the Interior Decorator. We hypothesize that this is due to the word choice of *Story* Editor because stories are often associated with creative writing, which does not have the level of objectivity necessary for the reasoning tasks in the TRIP benchmark.

5 Discussion of Results

In this section, we interpret the trends in the results we reported in the previous section, and we discuss the significance of our results in the field of commonsense reasoning. Overall, our results identify model selection and few-shot prompting as the most successful strategies to improve performance on the TRIP benchmark.

Conflict Types in 4 Few-Shot Examples	Accuracy ↑	Consistency ↑	Verifiability ↑
4 explicit vs. 0 implicit	50.00%	31.69%	11.27%
3 explicit vs. 1 implicit	55.63%	38.73%	16.90%
2 explicit vs. 2 implicit	54.23%	38.03%	11.97%
1 explicit vs. 3 implicit	59.15%	33.10%	9.86%
0 explicit vs. 4 implicit	66.20%	38.73%	8.45%

Table 8: Performance of Mistral-7B-Instruct-v0.3 with Different Distributions of Explicit vs. Implicit Conflicts in the Few-shot Examples

Conflict Types in 4 Few-Shot Examples	Accuracy ↑	Consistency ↑	Verifiability ↑
4 explicit vs. 0 implicit	50.70%	35.21%	13.38%
3 explicit vs. 1 implicit	51.41%	38.03%	13.38%
2 explicit vs. 2 implicit	52.11%	34.51%	8.45%
1 explicit vs. 3 implicit	49.30%	27.46%	14.08%
0 explicit vs. 4 implicit	49.30%	28.17%	9.15%

Table 9: Performance of Mistral-7B-v0.3 with Different Distributions of Explicit vs. Implicit Conflicts in the Few-shot Examples

5.1 Transfer Learning Inference

The BERT model transferred from CE has increased the accuracy of the TRIP benchmark, but not the consistency and verifiability. Given this result, we infer that CE aligns better with accuracy than consistency and verifiability. This result also suggests that fine-tuning on CE enhanced the model's ability to handle complex, abstract reasoning but may have sacrificed the detailed and grounded understanding needed for low-level tasks.

The results indicate that the BERT model transferred from PIQA underperformed compared to the original BERT model on the TRIP benchmark. This outcome suggests that while the PIQA fine-tuning process improved the model's understanding of physical reasoning, it may have led to over-specialization or an imbalance in general reasoning capabilities required for TRIP. An additional potential explanation is that the physical states learned from PIQA may not be similar enough to the physical states in TRIP. Previous work (Jiang et al., 2023b) aims to transfer knowledge from TRIP to PIQA, which is opposite to our work, and also confirms that 66.7% of task participants in PIQA are unseen during training with TRIP dataset.

Overall, our findings highlight the challenges of transfer learning, where task-specific fine-tuning can inadvertently reduce performance on tasks with distinct or broader requirements. Our results may also support concerns that models trained on commonsense reasoning datasets have not learned generalizable reasoning skills, but rather learned spurious correlations for solving that specific end task.

5.2 Heuristic-Analytic Reasoning in Contextual Language Models

The results reveal several critical insights into the performance of compact models and the effectiveness of HAR. Despite their smaller size, Mistral-7B and LLaMA-3.1-8B-Instruct achieved competitive verifiability scores, outperforming BERT in aligning high-level decisions with evidence-based justifications. This demonstrates the potential of compact architectures when combined with structured reasoning frameworks.

A closer examination of the metrics highlights the trade-offs between accuracy and coherence. While larger models, such as LLaMA-65B and InstructGPT, excel in accuracy, their dependency on extensive computational resources raises questions about their scalability. In contrast, the improved coherence metrics of compact models suggest a promising direction for resource-efficient reasoning.

The disparity between accuracy and coherence metrics also underscores the importance of multi-tiered evaluation. Models achieving high accuracy but low consistency or verifiability, such as BERT, may lack the reasoning coherence required for critical

Demonstration Set	Accuracy (%)	Consistency (%)	Verifiability (%)
Initial Demos #1	61.97	40.14	27.46
Modified Demos #1	62.96	40.14	29.57
Δ	+0.99	0.00	+2.11
Initial Demos #2	**70.42**	**50.00**	18.31
Modified Demos #2	66.2	49.29	21.26
Δ	-4.22	-0.71	+2.95

Table 10: Changes of Metrics without/with object replacement for Mistral-7B-Instruct Model, as evaluated on two different sets of initial demos.

Demonstration Set	Accuracy (%)	Consistency (%)	Verifiability (%)
Initial Demos #1	53.59	31.76	23.31
Modified Demos #1	52.11	31.06	25.35
Δ	-1.48	-0.07	+2.04
Initial Demos #2	52.11	32.39	14.08
Modified Demos #2	52.11	30.99	21.83
Δ	0.00	-1.40	+7.75

Table 11: Changes of Metrics with objects replacement for Mistral-7B Model, as evaluated on two different sets of initial demos.

applications. HAR addresses this gap by ensuring that heuristic decisions are consistently supported by analytic rationalizations, thereby enhancing trustworthiness.

In summary, our findings advocate for the continued exploration of HAR and compact models in reasoning tasks. Future research could extend this work by refining HAR strategies, exploring new datasets, and developing more efficient prompting mechanisms to further bridge the gap between compact and large models in commonsense reasoning tasks.

5.3 Targeted Prompts Improve Performance

Through the use of targeted prompting techniques, we were able to significantly improve upon the performance the Mistral models from our previous approach. Impressively, some trials saw up to 10% improvements in the accuracy and consistency metrics. The most significant performance improvements resulted from selecting a strong set of few-shot demonstrations that comprehensively portrayed the reasoning tasks to the model. In fact, carefully curated sets of 4 demonstrations frequently outperformed larger sets of 6 demonstrations.

However, our ability to improve performance by changing the few-shot examples also reflects the model's high sensitivity to the exact content and phrasing of the input prompt. In fact, we saw performance changes of up to 4% by simply changing the objects used in particular few-shot examples, even though these changes had no conceptual effect on the logic or plausibility of the story. In future work, we hypothesize that we could reduce this variance by fine-tuning the Mistral model on the PIQA dataset, similar to how we fine-tuned BERT on PIQA in Approach 1 of our project.

We also observed improved performance after implementing role-play prompting. By assigning the model to an Interior Decorator persona that is traditionally strong in physical state understanding, the model's performance on the verfiability metric (for physical state detection) improved.

Overall, this approach demonstrates the ability of carefully crafted prompts to strengthen the reasoning skills of LLMs. Successful prompts should frame the model's task with roles that have strong lower-level reasoning skills, and should utilize a set of diverse and informative few-shot examples that comprehensively represent the input space.

6 Conclusion

In this study, we explore three approaches to enhance reasoning performance on the TRIP benchmark: transfer learning, model selection, and ad-

vanced prompting techniques. Through comprehensive evaluations, we find that the Mistral-7B-Instruct-v0.3 model achieves the best performance, demonstrating the importance of both instruction fine-tuning and effective demonstration selection in improving reasoning coherence.

Our experiments reveal that carefully curated few-shot examples and targeted object replacements can significantly influence model performance across metrics like accuracy, consistency, and verifiability. However, the sensitivity of models to demonstration content emphasizes the need for continued research into optimizing input configurations.

Our future work could involve implementing the attention visualization from (Zhang et al., 2023) to gain deeper insights into reasoning processes under different prompting strategies. By visualizing how models process and adapt to varying inputs, we aim to better understand their decisions and further optimize prompting strategies.

7 Division of Work

Tiffany:

- **Implemented Code:**
  - **Model Selection:** Implemented Mistral-7B, Mistral-7B-Instruct, & Llama 3.1-8B
  - **Role-play Prompting:** Came up with our 2 roles and implemented role-play prompting
  - **Measuring Variance Caused By ICL Samples:** Updated HAR repository to create few-shot demos from a list of datapoint IDs specified via command line
- **Ran Experiments**:
  - **Model Selection:** Ran all of our 60 trials (10 trials for each of our 6 evaluated models) to establish the confidence interval for each model's performance
  - **Role-play Prompting:** Ran all trials for role-playing experiments
  - **Demo Selection:** Ran experiments on both Mistral models to measure performance changes as the distribution of conflict types varies
  - **Number of Few-shot Examples:** Ran experiments on Mistral-7B-v0.3 to measure how performance changes with different numbers of few-shot examples
- **Contributed to Deliverables:** Wrote Introduction and Related Work sections of report and presentation. Also wrote sections 4.43 and 5.3 about prompting techniques.
- **Contributed to Project Planning:** Came up with our project idea and our 3 approaches

Junkuan:

- **Implemented Code:**
  - **Explicit &Implicit demonstration** Implemented code to automatically select explicit and implicit demonstrations based on command line arguments.
- **Ran experiments:**
  - **Explicit & implicit demonstration** Tested how does the distribution of explicit & implicit demonstration influence the performance of llama-3.1-8B and llama-3.1-8B-instruct.
- **Contributed to Deliverables:**
  - **Presentation and Report:** Reported Model Selection part in the presentation. Wrote model selection part and explicit & implicit demonstrations part in the paper.

Yuting:

- **Implemented Code:**
  - **Transfer Learning** Finetuning BERT model with CE and PIQA and evaluation on TRIP.
- **Ran experiments:**
  - **Transfer Learning:** Ran BERT with different hyperparameters on CE, PIQA and TRIP training data and evaluation on TRIP test data.
- **Contributed to Deliverables:** Wrote Transfer Learning sections of report and presentation. Also wrote prompting with different number of demonstrations and readme of GiHub repo.

Xiyuan:

- **Implemented Code:**

- **Model Selection:** Adjust code for our environment and Llama-2 models; Implemented Llama 3.1-8B-Instruct
- **Variance Due To Demo Selection**: Implemented the replacements of the objects in the default 4 demos and the optimal demo set.

- **Ran experiments:**
  - **Model Evaluation:** Evaluate Llama-2-7B, Llama-2-13B and Llama-3.1-8B-Instruct with default 4 demos and balanced 6 demos.
  - **Evaluation demos with different objects:** Evaluate the performances of Mistral-7B-Instruct-v0.3 and Mistral-7B-v0.3 with (modified) Demo #1.
  - **Evaluation different demo set :** Evaluate the performances of Mistral-7B-Instruct-v0.3 with different demos specified by their IDs.
  - **Number of Few-shot Examples:** Ran experiments on Mistral-7B-Instruct-v0.3 to measure how performance changes with different numbers of few-shot examples.

- **Contributed to Deliverables:** Define objectives, organize workflows, and align deliverables with the project's goals.
  - Responsible for Experiments part in the report and presentation.

8 Link to Codebase

Our code is available at
https://github.com/tpari/TRIP_Team26/

Acknowledgments

Thank you to Joyce Chai and Martin Ma, both of whom we met with at office hours to solidify a project plan that was sufficiently interesting and advanced to demonstrate the depth of our NLP knowledge.

References

Berk Atil, Alexa Chittams, Liseng Fu, Ferhan Ture, Lixinyu Xu, and Breck Baldwin. 2024. Llm stability: A detailed analysis with some surprises. *Preprint*, arXiv:2408.04667.

Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the dangers of stochastic parrots: Can language models be too big? . In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, page 610–623, New York, NY, USA. Association for Computing Machinery.

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. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4762–4779, Florence, Italy. Association for Computational Linguistics.

Stevo Bozinovski. 2020. Reminder of the first paper on transfer learning in neural networks, 1976. *Informatica (Slovenia)*, 44(3).

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

Michael Chen, Mike D'Arcy, Alisa Liu, Jared Fernandez, and Doug Downey. 2019. CODAH: An adversarially-authored question answering dataset for common sense. In *Proceedings of the 3rd Workshop on Evaluating Vector Space Representations for NLP*, pages 63–69, Minneapolis, USA. 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. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Abhimanyu Dubey and et al. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Tyna Eloundou, Sam Manning, Pamela Mishkin, and Daniel Rock. 2023. Gpts are gpts: An early look at the labor market impact potential of large language models. *Preprint*, arXiv:2303.10130.

Suchin Gururangan, Swabha Swayamdipta, Omer Levy, Roy Schwartz, Samuel Bowman, and Noah A. Smith. 2018. Annotation artifacts in natural language inference data. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 107–112, New Orleans, Louisiana. Association for Computational Linguistics.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2020. Deberta: Decoding-enhanced BERT with disentangled attention. *CoRR*, abs/2006.03654.

Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, and Ting Liu. 2024. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Transactions on Information Systems*.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023a. Mistral 7b. *Preprint*, arXiv:2310.06825.

Yifan Jiang, Filip Ilievski, and Kaixin Ma. 2023b. Transferring procedural knowledge across commonsense tasks. *Preprint*, arXiv:2304.13867.

Jean Kaddour, Joshua Harris, Maximilian Mozes, Herbie Bradley, Roberta Raileanu, and Robert McHardy. 2023. Challenges and applications of large language models. *Preprint*, arXiv:2307.10169.

Aobo Kong, Shiwan Zhao, Hao Chen, Qicheng Li, Yong Qin, Ruiqi Sun, Xin Zhou, Enzhi Wang, and Xiaohang Dong. 2023. Better zero-shot reasoning with role-play prompting. *arXiv preprint arXiv:2308.07702*.

Lige Leng. 2024. Challenge, integration, and change: Chatgpt and future anatomical education. *Medical Education Online*, 29.

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. *CoRR*, abs/1907.11692.

Kaixin Ma, Filip Ilievski, Jonathan Francis, Eric Nyberg, and Alessandro Oltramari. 2022. Coalescing global and local information for procedural text understanding. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 1534–1545, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

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.

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, San Diego, California. Association for Computational Linguistics.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. *Preprint*, arXiv:2203.02155.

Kyle Richardson, Ronen Tamari, Oren Sultan, Reut Tsarfaty, Dafna Shahaf, and Ashish Sabharwal. 2022. Breakpoint transformers for modeling and tracking intermediate beliefs. *Preprint*, arXiv:2211.07950.

Maarten Sap, Ronan LeBras, 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. *Preprint*, arXiv:1811.00146.

Siqi Shen, Lajanugen Logeswaran, Moontae Lee, Honglak Lee, Soujanya Poria, and Rada Mihalcea. 2024. Understanding the capabilities and limitations of large language models for cultural commonsense. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 5668–5680, Mexico City, Mexico. Association for Computational Linguistics.

Robyn Speer, Joshua Chin, and Catherine Havasi. 2018. Conceptnet 5.5: An open multilingual graph of general knowledge. *Preprint*, arXiv:1612.03975.

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.

Niket Tandon, Bhavana Dalvi Mishra, Joel Grus, Wen tau Yih, Antoine Bosselut, and Peter Clark. 2018. Reasoning about actions and state changes by injecting commonsense knowledge. *Preprint*, arXiv:1808.10012.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2023. Self-consistency improves chain of thought reasoning in language models. *Preprint*, arXiv:2203.11171.

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.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William W. Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. 2018. Hotpotqa: A dataset for diverse, explainable multi-hop question answering. *Preprint*, arXiv:1809.09600.

Chen Zhang and Joyce Chai. 2010. Towards conversation entailment: An empirical investigation. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 756–766, Cambridge, MA. Association for Computational Linguistics.

Shengyu Zhang, Linfeng Dong, Xiaoya Li, Sen Zhang, Xiaofei Sun, Shuhe Wang, Jiwei Li, Runyi Hu, Tianwei Zhang, Fei Wu, and Guoyin Wang. 2024. Instruction tuning for large language models: A survey. *Preprint*, arXiv:2308.10792.

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. *Preprint*, arXiv:2310.18364.

Barret Zoph, Golnaz Ghiasi, Tsung-Yi Lin, and and Ekin D. Cubuk and Quoc V. Le Yin Cui, and Hanxiao Liu. 2020. Rethinking pre-training and self-training. page 3833–3845. Advances in Neural Information Processing Systems.