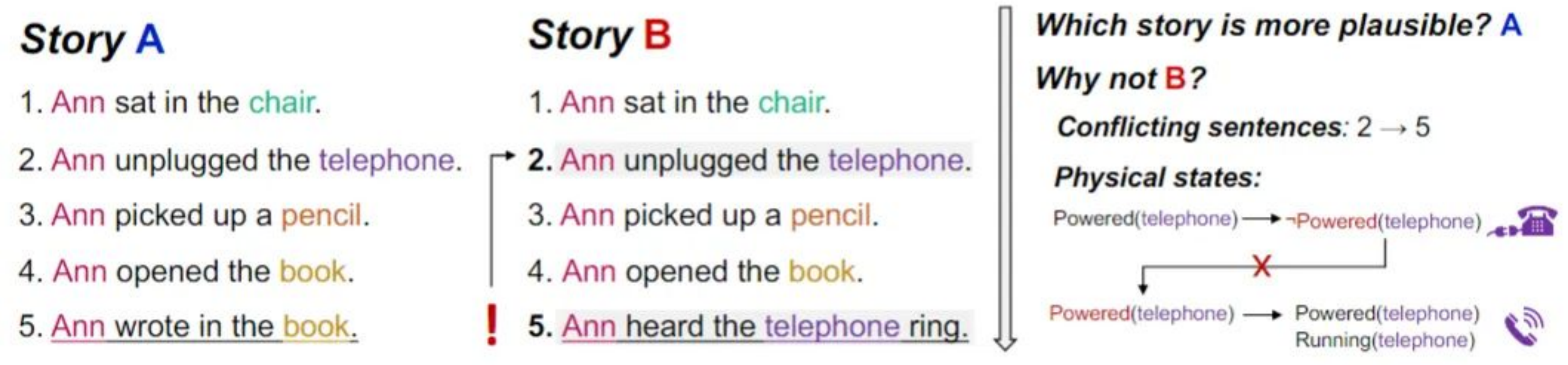
Approaches to Improve Performance of LLMs on TRIP Data Set

Team 24: Ishita Deshmukh, Carlos Figueredo, Hanning Li, Fei Wu

Introduction

- Tiered Reasoning for Intuitive Physics (TRIP) Dataset:
    - NLP dataset from the SLED Lab at the University of Michigan
    - Focused on multi-layered physical commonsense reasoning
- Problem:
    - Current baseline systems only achieve 10% accuracy with proper justification
    - Goal is to achieve greater accuracy with correct justification



Story A

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

Story B

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

***Which story is more plausible?* A**

Why not* B*?

Conflicting sentences: 2 → 5

Physical states:

!
Powered(telephone) → ¬Powered(telephone)
X
Powered(telephone) → Powered(telephone)
Running(telephone)

[1]

Relation to Previous Work

- Models
  - BERT
  - RoBERTa
  - DeBERTa
- Evaluation metrics
  - Accuracy — correct story identification
  - Consistency — correct conflict identification
  - Verifiability — correct physical state changes
- Other datasets that have been explored
  - ROCSTories
  - PIQA
  - SPARTQA
  - ProPara

Methods

- Methods 1-3: Use different models on the TRIP dataset
    - ALBERT
    - Mistral-7b
    - GPT-2 XL
- Method 4: Answer Set Programming

Method #4: Using ASP

Answer Set Programming: declarative programming paradigm designed for solving complex search and optimization problems

Premises:

1) “Socrates is a human”
2) “All humans are mortal”

Conclusion:

“Socrates is ……….”

```
% Facts
human(socrates).

% Rules
mortal(X) :- human(X).

% Query
#show mortal(socrates).
```

Note: Initial approach was to use a solver-augmented LLM idea to solve TRIP’s tasks, but this approach failed.

Method #4 Updated: Prompting using ASP

Does ASP’s rich logical semantics add value to the LLM’s reasoning?

EXPERIMENT

Task: Select which of the following 2 stories is more plausible. 1 or 2? (1st Task of TRIP)

Size: 100 samples from Test split

Metric: Accuracy

Strategies: Zero-shot, Few-shot, ASP Few-shot

MISTRAL
AI_

`Mistral-7B-Instruct-v0.3`

Preliminary Results: Methods #1-3

Implementation of these methods is still ongoing

	Accuracy	Consistency	Verifiability
GPT-2 XL	3.6%	3.1%	0.9%
ALBERT	72.6%	5.7%	2.0%

Results: Method #4

Does ASP’s rich logical semantics add value to the LLM’s reasoning?

- Short Answer: No.
- However, ...

Prompting Strategy	Accuracy
Zero-shot	58%
Few-shot	68%
ASP Few-shot	62%

ASP Few-shot can generate implicit rules:

```
implausible(C) :- collection(C), sequence(C, T1, smash_radio),
                      sequence(C, T2, switch_on_radio), T2 > T1.
```

Plan

- Methods #1-3:
  - Complete experiments with Mistral-7b and GPT-2 XL
  - Experiment with different loss configurations
- Method #4:
  - Evaluate TRIP's 2nd Task. Which sentences make the story implausible?
  - Use **consistency** metric.
  - How? Process ASP Few-shot's generated answer.

References

[1] Shane Storks, Qiaozi Gao, Yichi Zhang, and Joyce Chai. 2021. Tiered reasoning for intuitive physics: Toward verifiable commonsense language understanding. *arXiv:2109.04947*.

Approaches to Improve Performance of LLMs on TRIP Data Set

Ishita Deshmukh, Carlos Figueredo, Hanning Li, Fei Wu