Tiered Reasoning for Intuitive Physics:

Toward Verifiable Commonsense Language Understanding

Shane Storks, Qiaozi Gao, Yichi Zhang, & Joyce Chai

(he/him)

Situated Language and Embodied Dialogue (SLED)

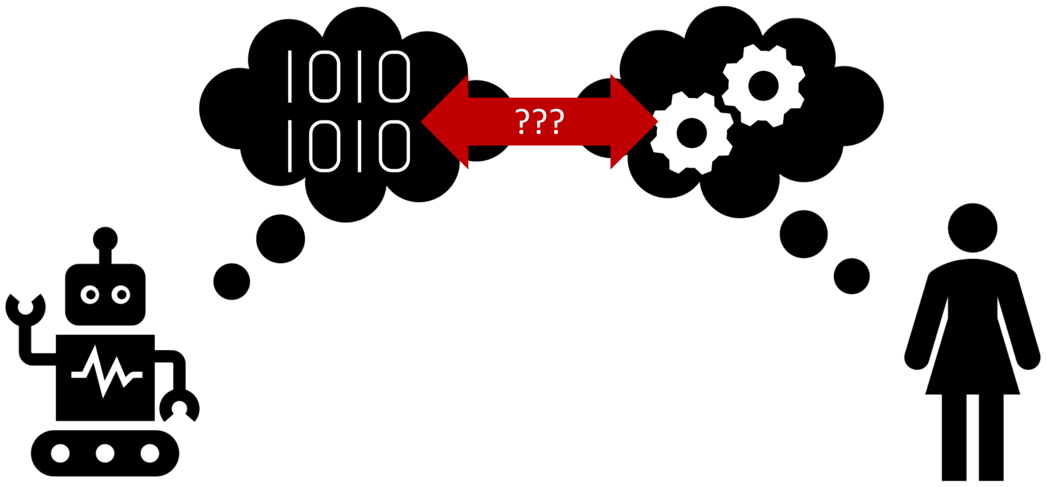
University of Michigan, Computer Science and Engineering Division

sstorks@umich.edu

Findings of EMNLP 2021 Long Paper

Motivation

- Large-scale, pre-trained LMs are nearing and surpassing human performance on many language understanding tasks!
- It remains unclear whether the problems are *truly solved* 🧐
    - Lack of interpretability
    - Data bias
- How can we *verify* the reasoning of large LMs?



Tiered Reasoning for Intuitive Physics (TRIP)

- New dataset providing traces of a multi-tiered, human-annotated reasoning process:
  - Low-level, concrete physical states
  - High-level end task of plausibility classification

Tiered Reasoning for Intuitive Physics (TRIP)

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

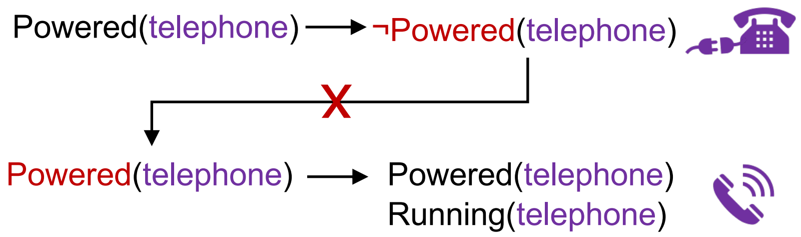
1. Ann sat in the chair.
2. Ann unplugged the telephone.
3. Ann picked up a pencil.
4. Ann opened the book.
!
5. Ann heard the telephone ring.

Which story is more plausible? **A**

Why not **B*****?***

Conflicting sentences*:* 2 → 5

Physical states:



Powered(telephone) → ¬Powered(telephone) ☎

Powered(telephone) → Powered(telephone) Running(telephone)

Data Statistics

- **675 plausible stories**
  - 370 train, 152 validation, 153 test
- **1476 implausible stories**
  - 802 train, 323 validation, 351 test
- 6 everyday environments
  - `kitchen`, `bathroom`, `living room`, `garage`, `office`, `park`
- Vocabulary size (overall): 2126
  - 486 verbs, 781 nouns

Data Statistics

- Average of 1.2 conflicting sentence pairs per implausible story
- 36.6k labels of physical states
  - 18.8k train, 8.74k validation, 9.09k test
- 20 annotated attributes

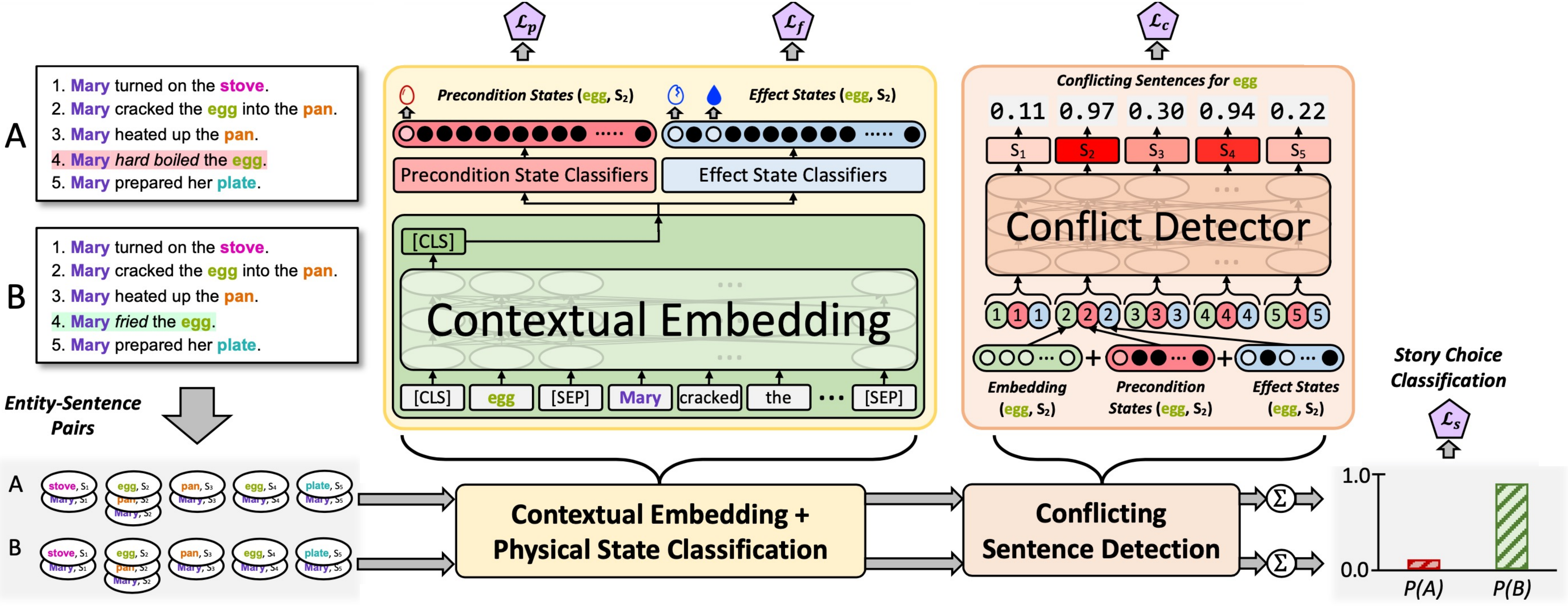
- ***Humans***
  1. Location
  2. Conscious
  3. Wearing
  4. Wet
  5. Hygiene

- ***Objects***
  1. Location
  2. Exist
  3. Clean
  4. Power
  5. Functional
  6. Pieces
  7. Wet
  8. Open
  9. Temperature
  10. Solid
  11. Contain
  12. Running
  13. Moveable
  14. Mixed
  15. Edible

Evaluation Metrics

Metric	Story Choice	Conflicting Sentences	Physical States
Accuracy	✓		
Consistency	✓	✓	
Verifiability	✓	✓	✓

Tiered Baseline



A

1. Mary turned on the stove.
2. Mary cracked the egg into the pan.
3. Mary heated up the pan.
4. Mary *hard boiled* the egg.
5. Mary prepared her plate.

B

1. Mary turned on the stove.
2. Mary cracked the egg into the pan.
3. Mary heated up the pan.
4. Mary *fried* the egg.
5. Mary prepared her plate.

$$\mathcal{L} = \lambda_p \mathcal{L}_p + \lambda_f \mathcal{L}_f + \lambda_c \mathcal{L}_c + \lambda_s \mathcal{L}_s$$

Loss Configuration	Model	Accuracy (%)	Consistency (%)	Verifiability (%)
--	random	47.8	11.3	0.0
All Losses	BERT	**78.3**	2.8	0.0
	RoBERTa	75.2	6.8	0.9
	DeBERTa	74.8	2.2	0.0
Omit Story Choice Loss $\mathcal{L}_s$	BERT	73.9	**28.0**	9.0
	RoBERTa	73.6	22.4	**10.6**
	DeBERTa	75.8	24.8	7.5
Omit Conflict Detection Loss $\mathcal{L}_c$	BERT	50.9	0.0	0.0
	RoBERTa	49.7	0.0	0.0
	DeBERTa	52.2	0.0	0.0
Omit State Classification Losses $\mathcal{L}_p$ and $\mathcal{L}_f$	BERT	75.2	17.4	0.0
	RoBERTa	71.4	2.5	0.0
	DeBERTa	72.4	9.6	0.0

All losses ⇒ low consistency & verifiability.

No end-task loss ⇒ better consistency & verifiability!

Conflict detection doesn't emerge naturally.

Physical states don't emerge naturally either.

Devlin, J., Chang, M., Lee, K., & Toutanova, K. (2019). BERT: Pre-Training of Deep Bidirectional Transformers for Language Understanding. NAACL HLT 2019.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., & Stoyanov, V. (2019). RoBERTa: A Robustly Optimized BERT Pretraining Approach. arXiv: 1907.11692.

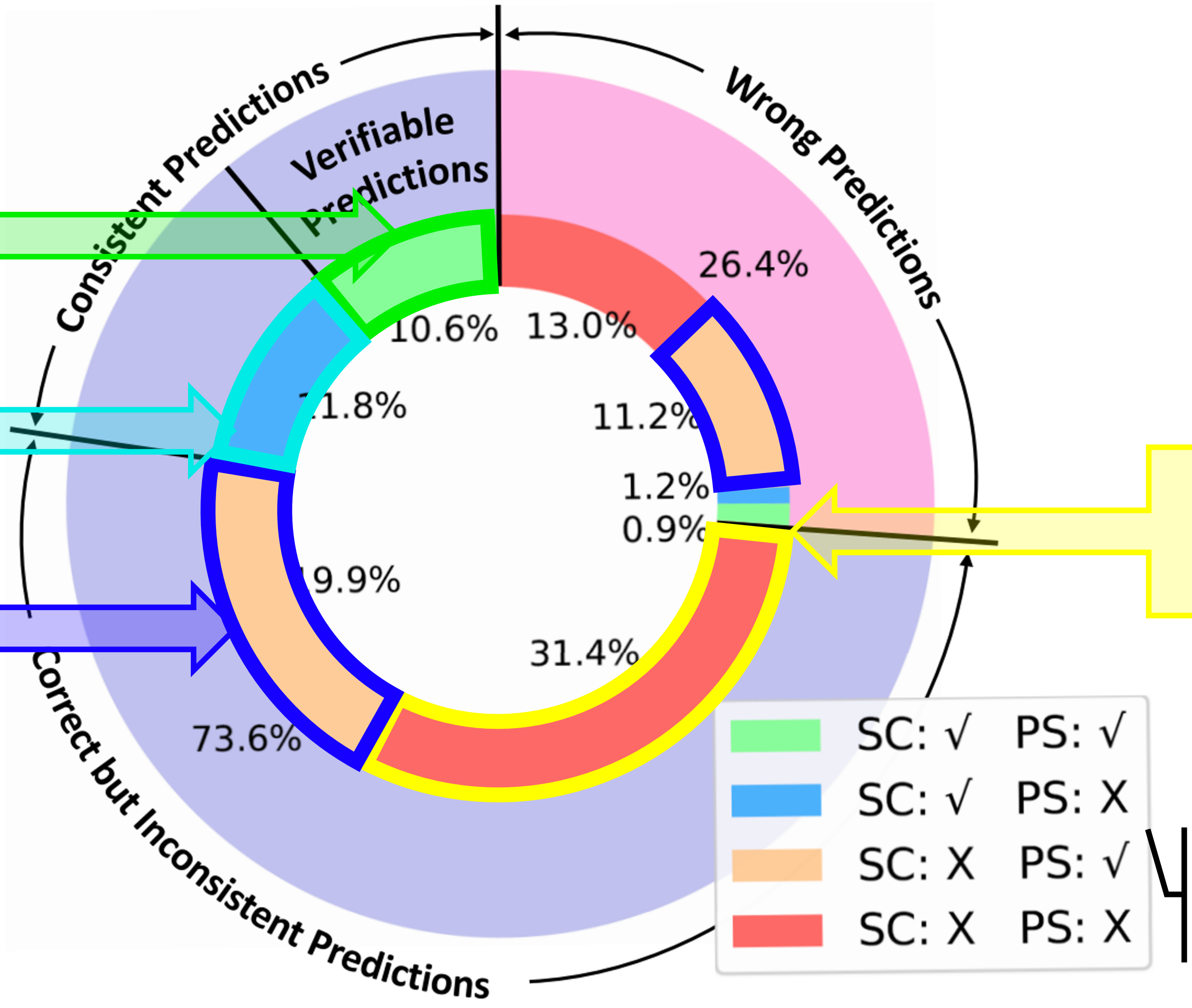
He, P., Liu, X., Gao, J., & Chen, W. (2021). DeBERTa: Decoding-enhanced BERT with Disentangled Attention. arXiv: 2006.03654.

Error Distribution

Correct and entirely verifiable!

Consistent but not verifiable!

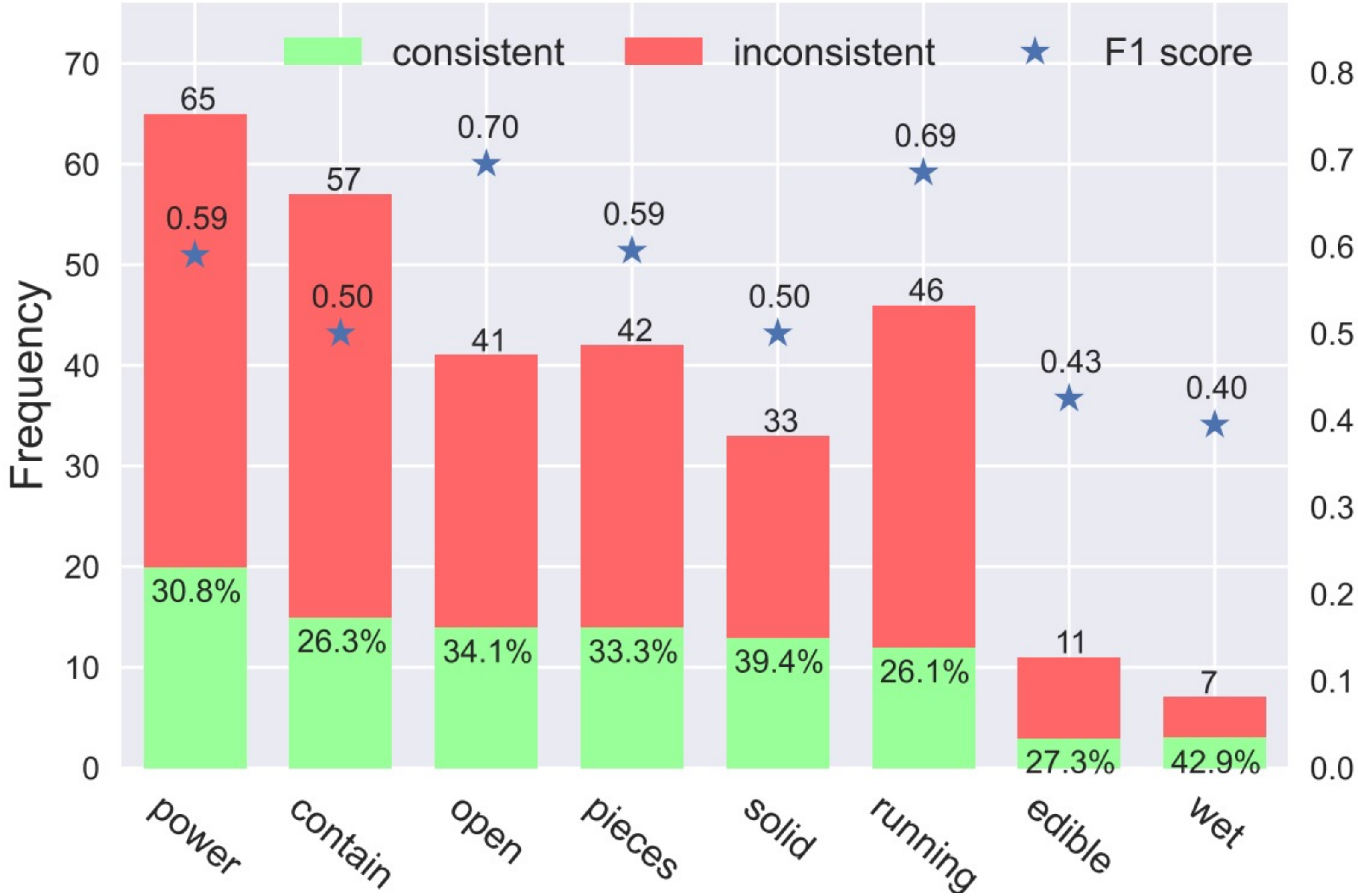
Correct states, but unsuccessful conflict detection. 🤔



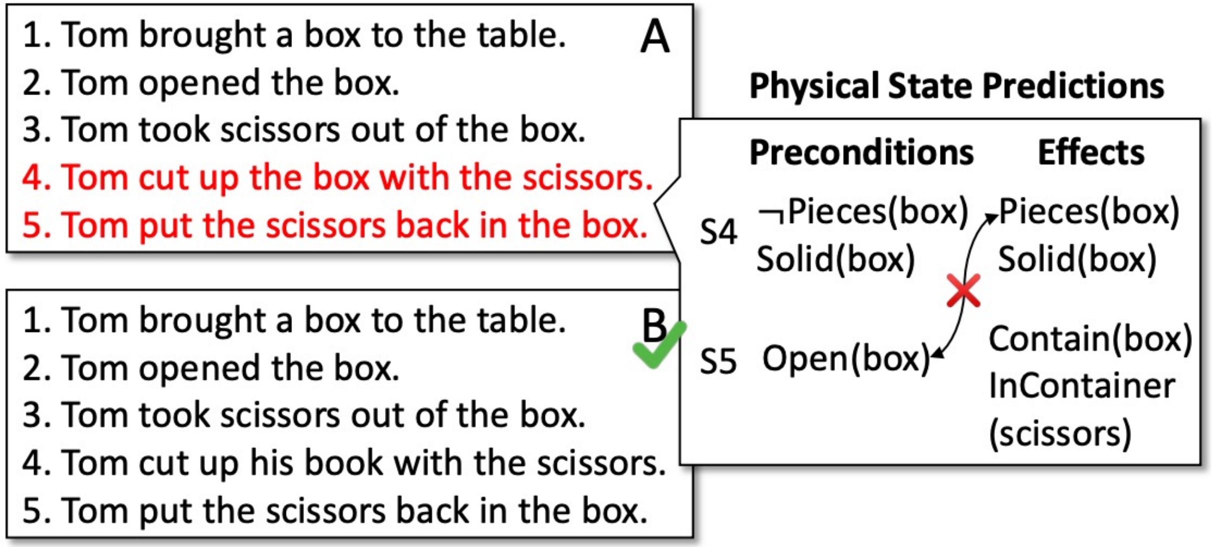
Correct, but entirely unverifiable!

SC: sentence conflict
PS: physical states

Utility of Attributes



Sample System Outputs



1. Tom brought a box to the table.
2. Tom opened the box.
3. Tom took scissors out of the box.
4. Tom cut up the box with the scissors.
5. Tom put the scissors back in the box.

1. Tom brought a box to the table.
2. Tom opened the box.
3. Tom took scissors out of the box.
4. Tom cut up his book with the scissors.
5. Tom put the scissors back in the box.

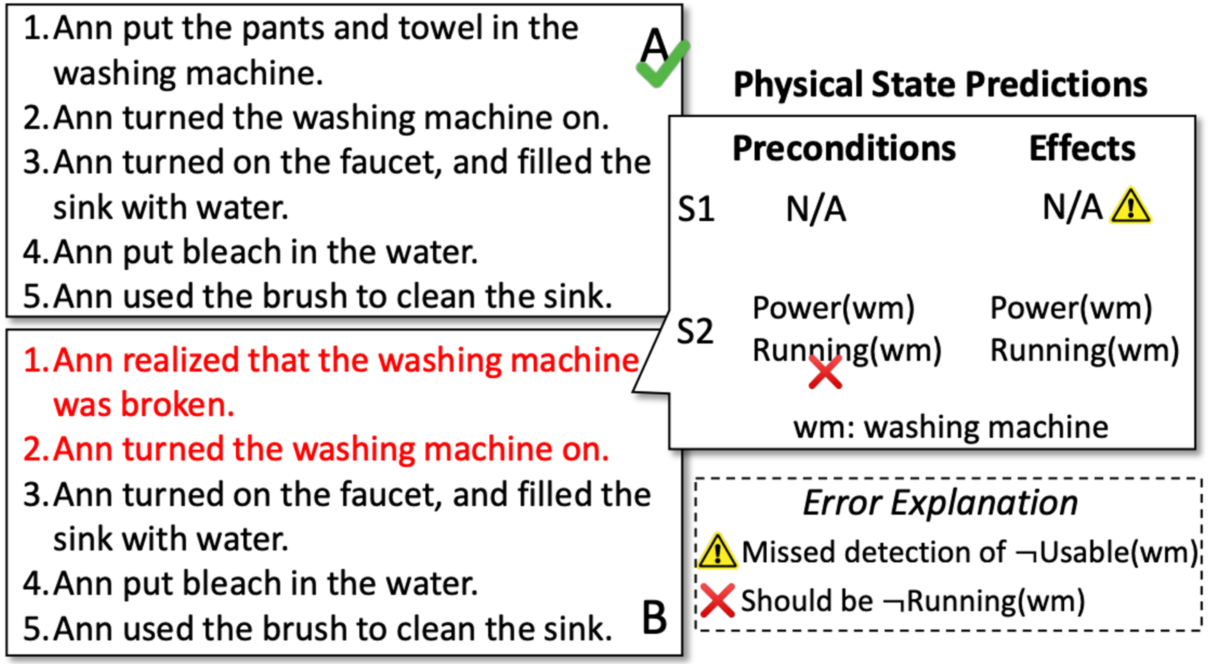
Physical State Predictions

	Preconditions	Effects

S4	¬Pieces(box)
	Solid(box)
		Pieces(box)
		Solid(box)

S5	Open(box) ✗	Contain(box) InContainer (scissors)

(a) A verifiable prediction.



1. Ann put the pants and towel in the washing machine.
2. Ann turned the washing machine on.
3. Ann turned on the faucet, and filled the sink with water.
4. Ann put bleach in the water.
5. Ann used the brush to clean the sink.

1. Ann realized that the washing machine was broken.
2. Ann turned the washing machine on.
3. Ann turned on the faucet, and filled the sink with water.
4. Ann put bleach in the water.
5. Ann used the brush to clean the sink.

Physical State Predictions

	Preconditions	Effects
S1	N/A	N/A ⚠

S2	Power(wm)
	Running(wm) ✗
		Power(wm)
		Running(wm)

wm: washing machine

Error Explanation

⚠ Missed detection of ¬Usable(wm)

✗ Should be ¬Running(wm)

(b) A consistent but not verifiable prediction.

Summary

1. TRIP, a **novel multi-tiered dataset** enabling training and evaluation of commonsense reasoning verifiability in NLP models.

2. Large LMs **struggle to learn verifiable reasoning strategies** when trained as tiered, verifiable reasoning systems.

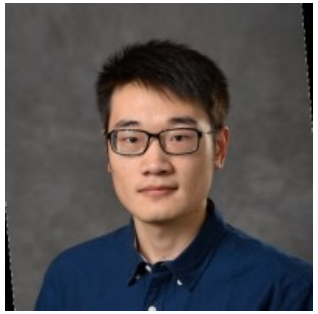
Summary

1. TRIP, a **novel multi-tiered dataset** enabling training and evaluation of commonsense reasoning verifiability in NLP models.

2. Large LMs **struggle to learn verifiable reasoning strategies** when trained as tiered, verifiable reasoning systems.

Acknowledgements

- **Advisor**: Joyce Chai
- **Collaborators**:
  - Qiaozi Gao
  - Yichi Zhang
- **Undergraduate assistants**:
  - Bri Epstein
  - Haoyi Qiu



Thank you!