Tiered Reasoning for Intuitive Physics: Toward Verifiable Commonsense Language Understanding

Shane Storks, Qiaozi Gao, Yichi Zhang, & Joyce Chai
Computer Science and Engineering Division, University of Michigan

Introduction

In light of large, pre-trained language models (LMs) nearing and surpassing human performance on a breadth of language understanding tasks [1, 2, 3], we propose Tiered Reasoning for Intuitive Physics (TRIP), a more challenging evaluation targeting physical commonsense in a densely annotated, tiered reasoning setting:

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)

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

Baseline Approach

We propose a tiered architecture powered by large, pre-trained LMs and their contextual embeddings:

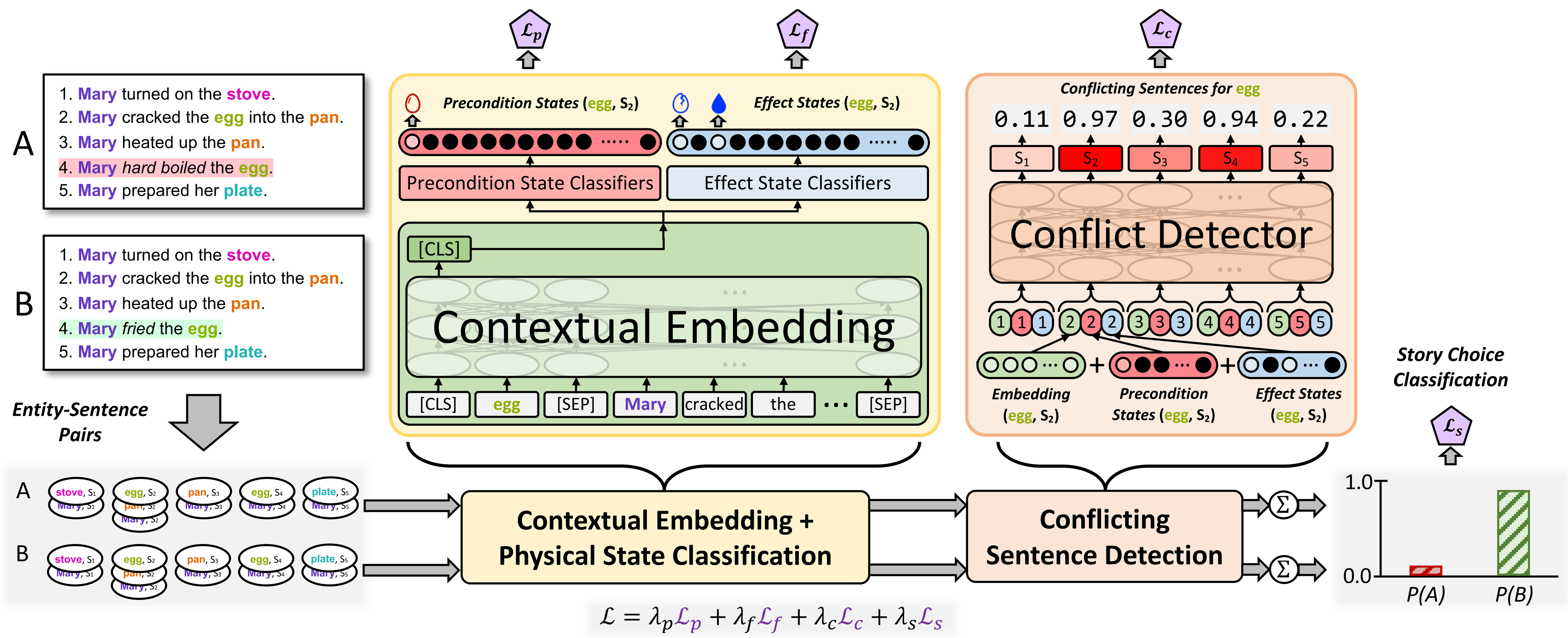

Figure 2: Proposed tiered reasoning system that generates contextual embeddings of entity-sentence pairs using a pre-trained LM, then uses them to jointly predict physical states, detect conflicting sentences, and identify the plausible story. The model is trained end-to-end by optimizing a weighted sum of cross-entropy loss functions $\mathcal{L}_p$ for precondition state classification, $\mathcal{L}_f$ for effect state classification, $\mathcal{L}_c$ for conflicting sentence detection, and $\mathcal{L}_s$ for story choice classification.

Baseline Results

We evaluate systems with three metrics:

1. **Accuracy**: requires story choice to be correct.
2. **Consistency**: additionally requires conflicting sentences in the implausible story to be correct.
3. **Verifiability**: additionally requires some physical states to be predicted for the conflicting sentences, and all predicted states must be correct.

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

Table 1: End and tiered task metrics for tiered classifiers on the validation set of TRIP trained on varied combinations of loss functions. Random baseline averaged over 10 runs.

Sample Outputs

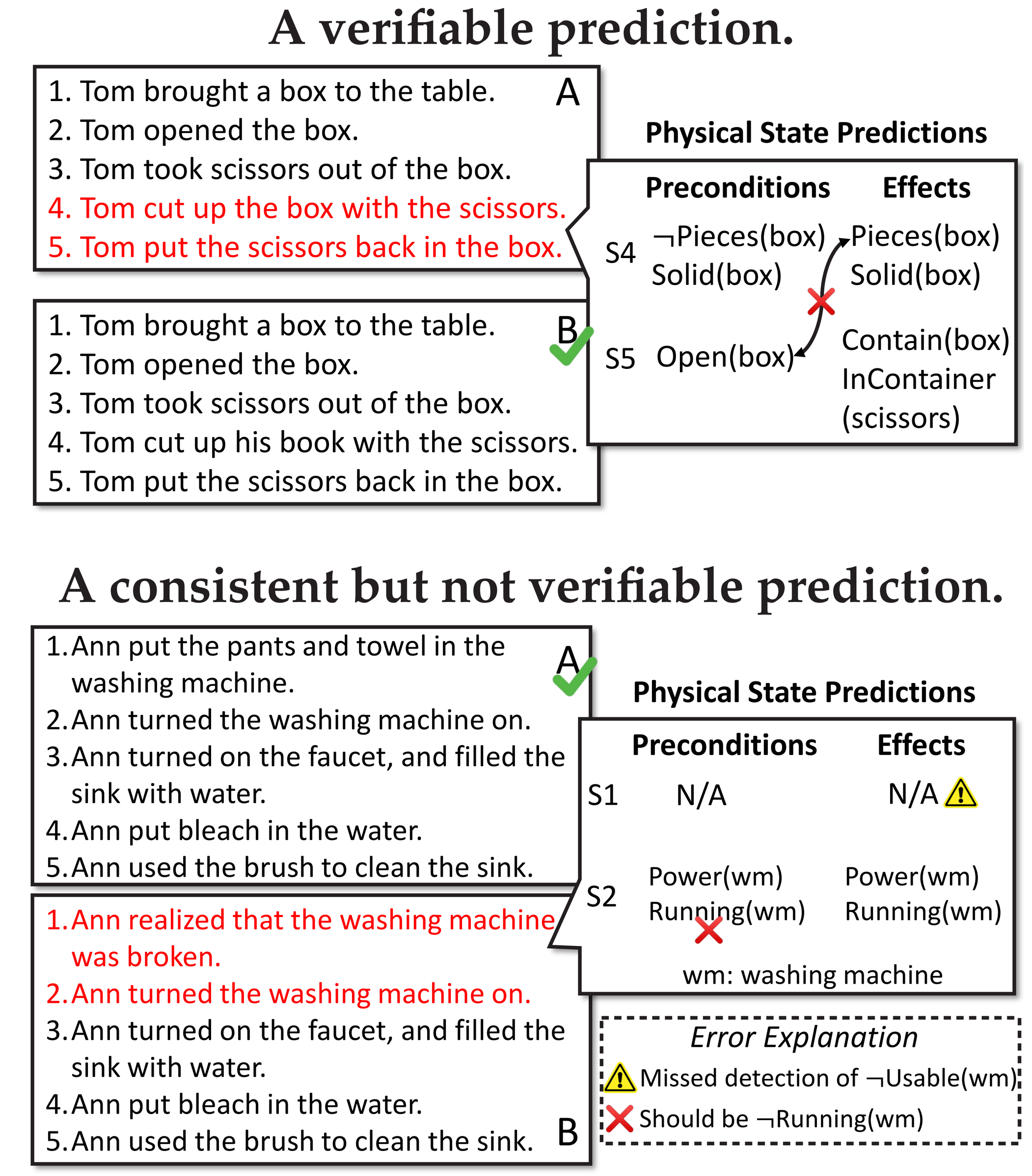

Figure 3: Sample outputs from the baseline system. The detected conflicting sentences are in red, and physical state predictions are shown on the right.

Reasoning Breakdown

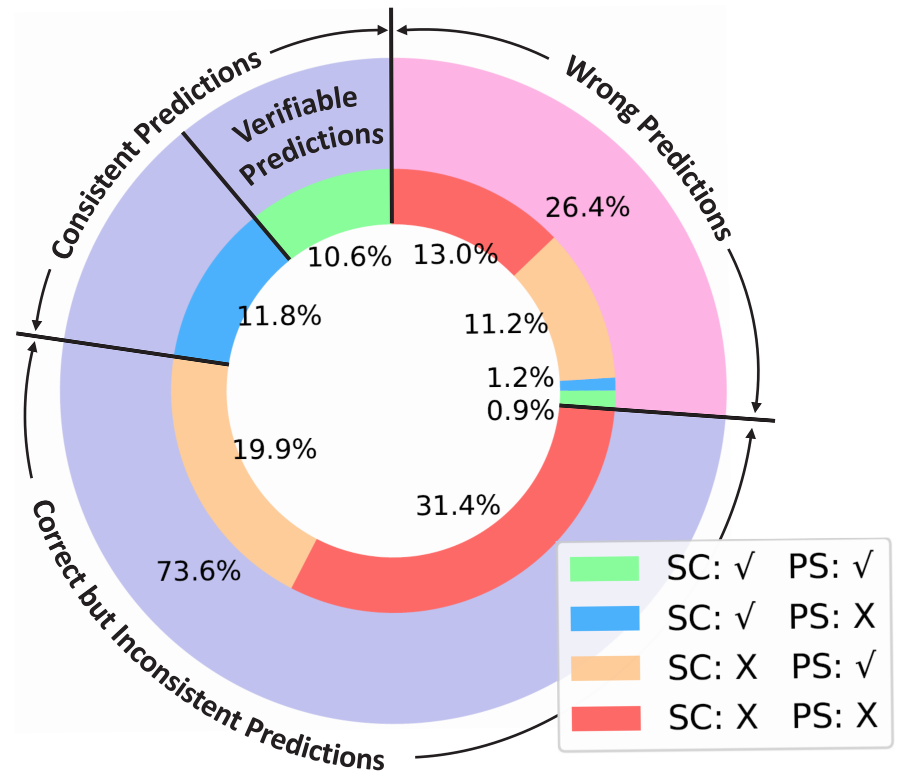

Figure 4: Distribution of RoBERTa successes and failures on TRIP. SC (sentence conflict) and PS (physical state) denote whether the predicted conflicting sentences or physical states are correct (✓) or not (×).

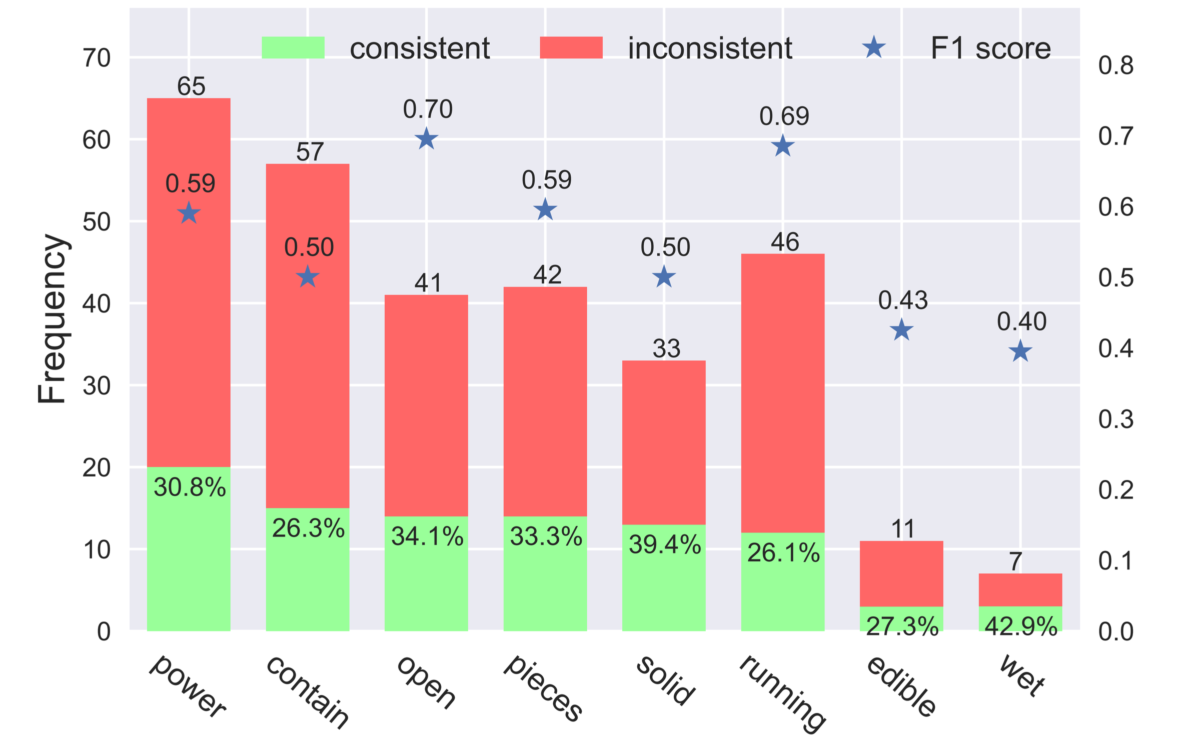

Figure 5: *Utility* of physical state predictions for selected attributes. Among correctly predicted physical states, bar regions indicate how many contribute to consistent end task predictions (i.e., with successfully detected conflicts). Blue stars indicate macro-F1 score of state prediction.

Conclusion

Our results show that *supervising large LMs based on high-level classification tasks in order to learn commonsense language understanding leads to inconsistent and unverifiable reasoning*. In order to solve tasks like these coherently, we should directly train systems to incorporate multiple types of lower-level evidence. Our work provides an important first step toward this goal and strong intuition for future progress.

References

[1] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. BERT: Pre-Training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of NAACL HLT 2019*, 2019.

[2] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov. RoBERTa: A Robustly Optimized BERT Pretraining Approach. *arXiv: 1907.11692*, 2019.

[3] P. He, X. Liu, J. Gao, and W. Chen. DeBERTa: Decoding-enhanced BERT with Disentangled Attention. *arXiv:2006.03654*, 2021.

Links

Scan the QR codes for related websites:

Preprint | Code | Lab Site