Toward Coherent Commonsense Language Understanding in Machines

Shane Storks

(he/him)

Situated Language and Embodied Dialogue

sstorks@umich.edu

EECS 692 (Advanced Artificial Intelligence) Guest Lecture

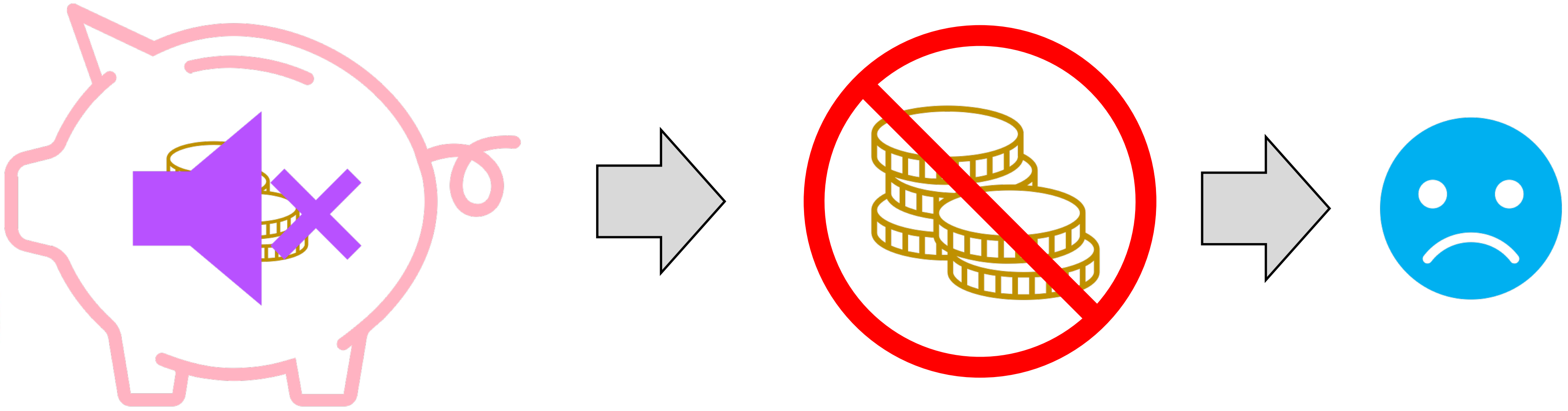
Outline

1. Introduction
2. Current Limitations
3. Assessing Coherence of Commonsense Reasoning
4. Learning Verifiable Commonsense Reasoning
5. Conclusion

Commonsense Reasoning

"Jack needed some money, so he went and shook his piggy bank. He was disappointed when it made no sound."

(dreamstime)



Minsky, M. (2000). Commonsense-based interfaces. In *Commun. ACM*, 43(8): p. 66-73.
Davis, E. & Marcus, G. (2015). Commonsense reasoning and commonsense knowledge in artificial intelligence. In *Commun. ACM*, 58(9): p. 92-103.

Then what is all this about?

New AI Model Exceeds Human Performance at Question Answering

(BecomingHuman.ai)

Dave Costenaro Follow

Nov 21, 2018 · 5 min read

AI models from Microsoft and Google already surpass human performance on the SuperGLUE language benchmark

Kyle Wiggers @Kyle_L_Wiggers January 6, 2021 11:04 AM

(The Machine)

AI, ML & DATA ENGINEERING

InfoQ Live (June 22nd) - Overcome Cloud and Serverless Security Challe

AI Models from Google and Microsoft Exceed Human Performance on Language Understanding Benchmark

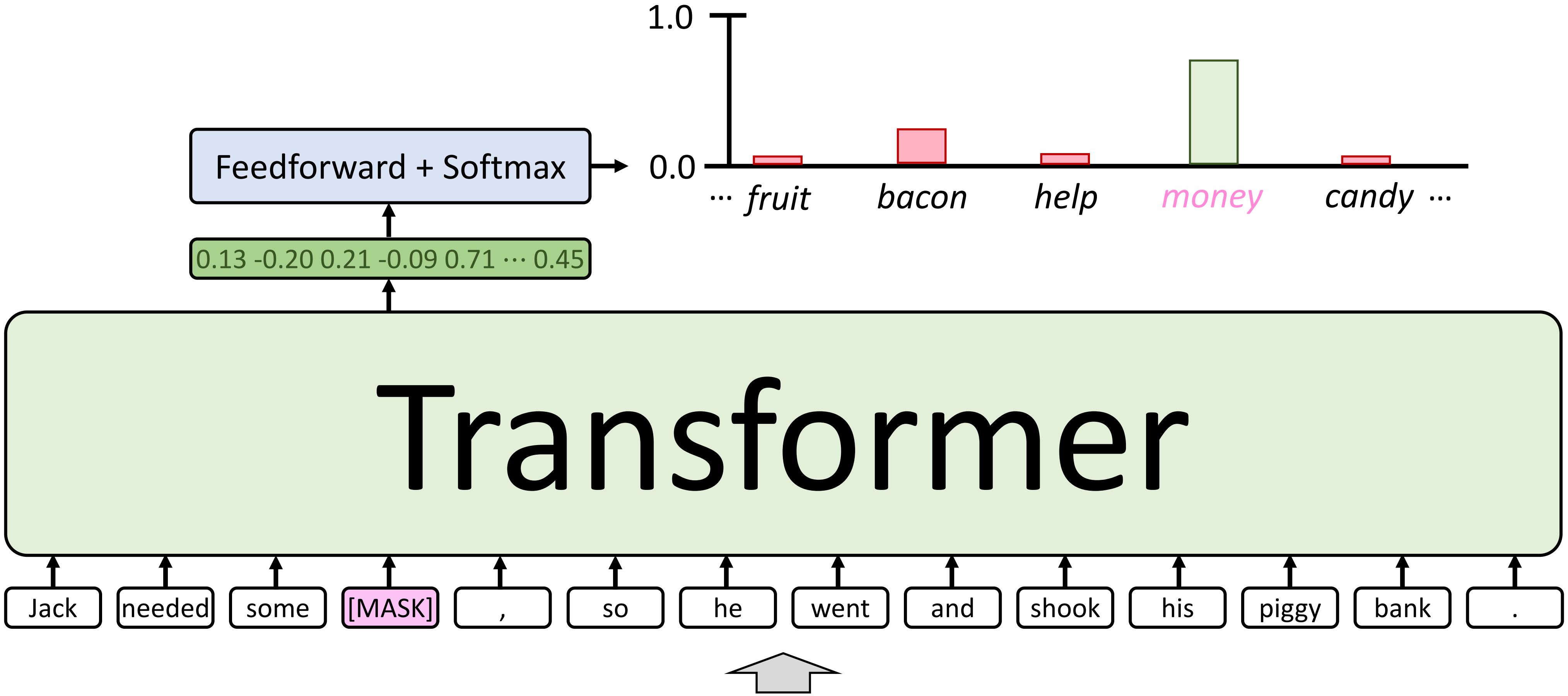
LIKE DISCUSS

(InfoQ)

JAN 12, 2021 • 3 MIN READ

Research teams from Google and Microsoft have recently developed natural language

Large, Pre-Trained Language Models (LMs)



"Jack needed some ***money****, so he went and shook his piggy bank."*

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-Training of Deep Bidirectional Transformers for Language Understanding. In NAACL HLT 2019.
Vaswani, A. et al. (2017). Attention is All you Need. In NIPS 30.

Large, Pre-Trained Language Models (LMs)

```
Q: What is your favorite animal?
A: My favorite animal is a dog.

Q: Why?
A: Because dogs are loyal and friendly.

Q: What are two reasons that a dog might be in a bad mood?
A: Two reasons that a dog might be in a bad mood are if it is
   hungry or if it is hot.

Q: How many eyes does a giraffe have?
A: A giraffe has two eyes.
```

Downstream Classification Tasks

Which sentence is most likely to fill in the blank?

It was a very hot summer day.

____

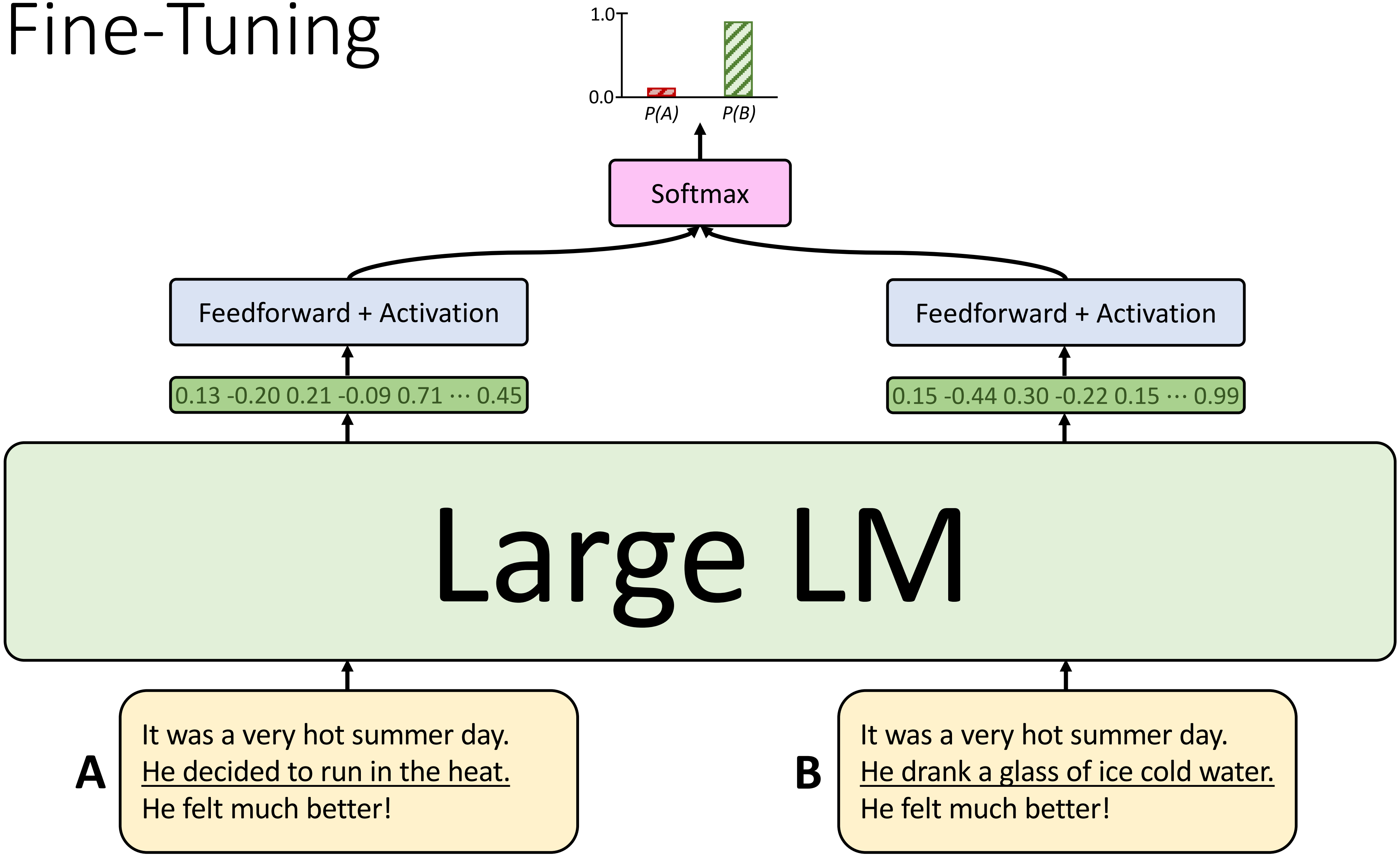
He felt much better!

He decided to run in the heat.

He drank a glass of ice cold water.

Bhagavatula, C., Le Bras, R., Malaviya, C. et al. (2020). Abductive commonsense reasoning. In ICLR 2020.

Fine-Tuning

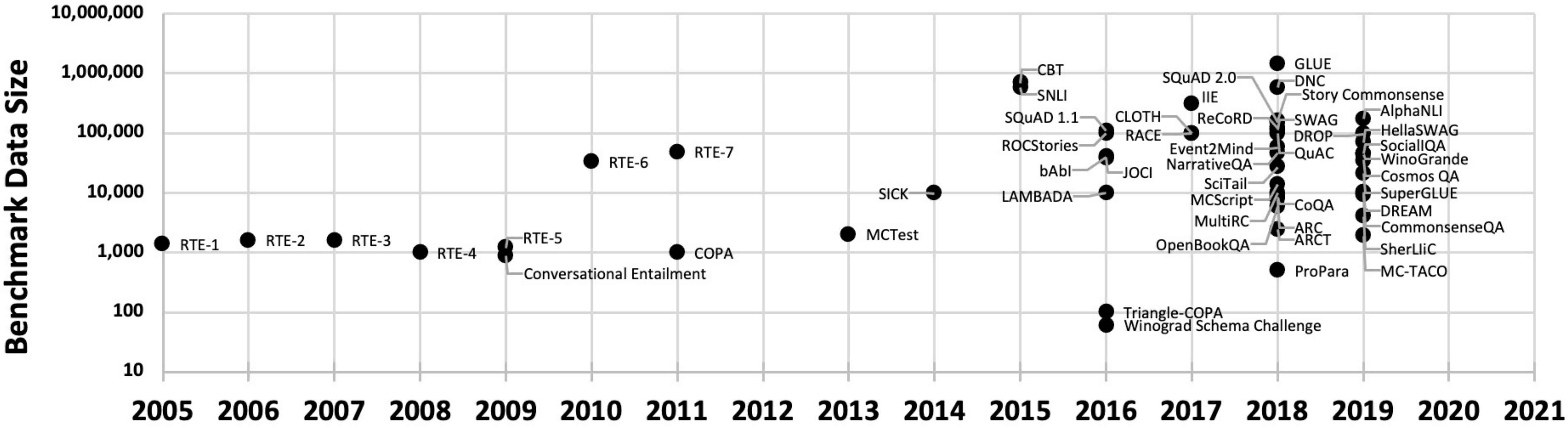

1.0
0.0
P(A)
P(B)
Softmax
Feedforward + Activation
Feedforward + Activation
0.13 -0.20 0.21 -0.09 0.71 ··· 0.45
0.15 -0.44 0.30 -0.22 0.15 ··· 0.99
Large LM
A
It was a very hot summer day.
He decided to run in the heat.
He felt much better!
B
It was a very hot summer day.
He drank a glass of ice cold water.
He felt much better!

Leaderboard Ranking

Rank	Submission	Created	Accuracy
1	**UNIMO** *UNIMO Team, Baidu NLP*	05/15/2021	**0.9118**
2	**DeBERTa** *Microsoft Dynamics 365 AI*	10/27/2020	0.8970
3	**anonymous**	04/22/2021	0.8783
4	**UNICORN** *Anonymous*	07/23/2020	0.8734
5	**anonymous** *ai2*	05/04/2021	0.8730

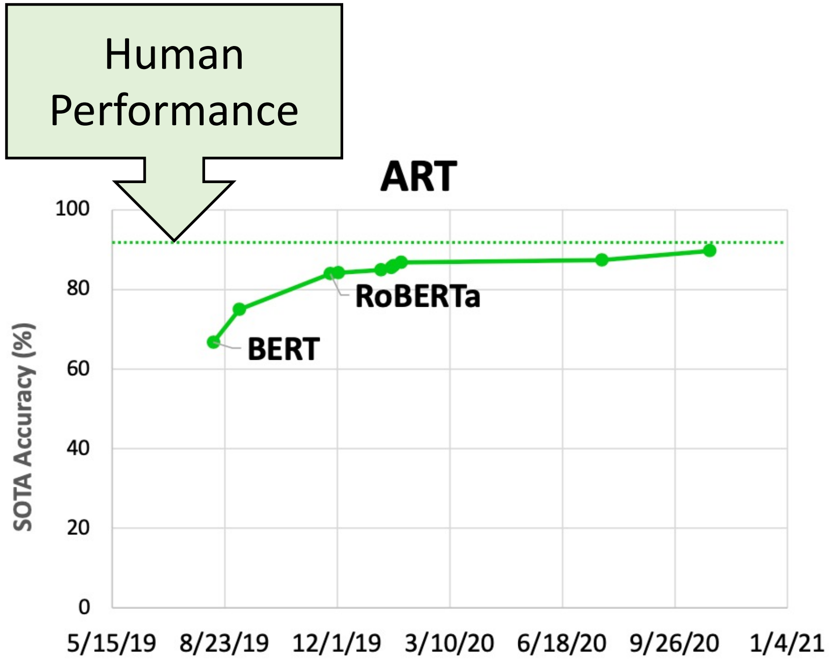
https://leaderboard.allenai.org/anli/submissions/public

Benchmark Datasets

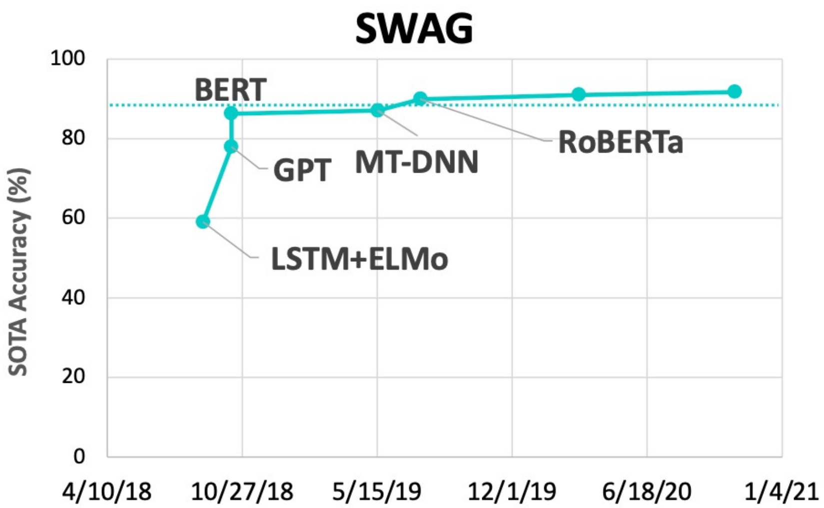


Storks, S., Gao, Q., & Chai, J.Y. (2020). Recent Advances in Natural Language Inference: A Survey of Benchmarks, Resources, and Approaches. arXiv: 1904.011672 [cs.CL].

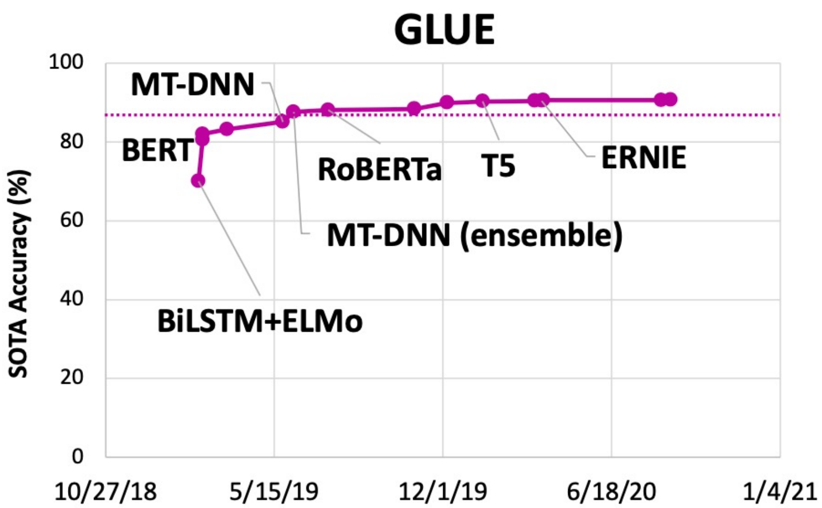
Human-Level Results



https://leaderboard.allenai.org/anli/submissions/public



https://leaderboard.allenai.org/swag/submissions/public



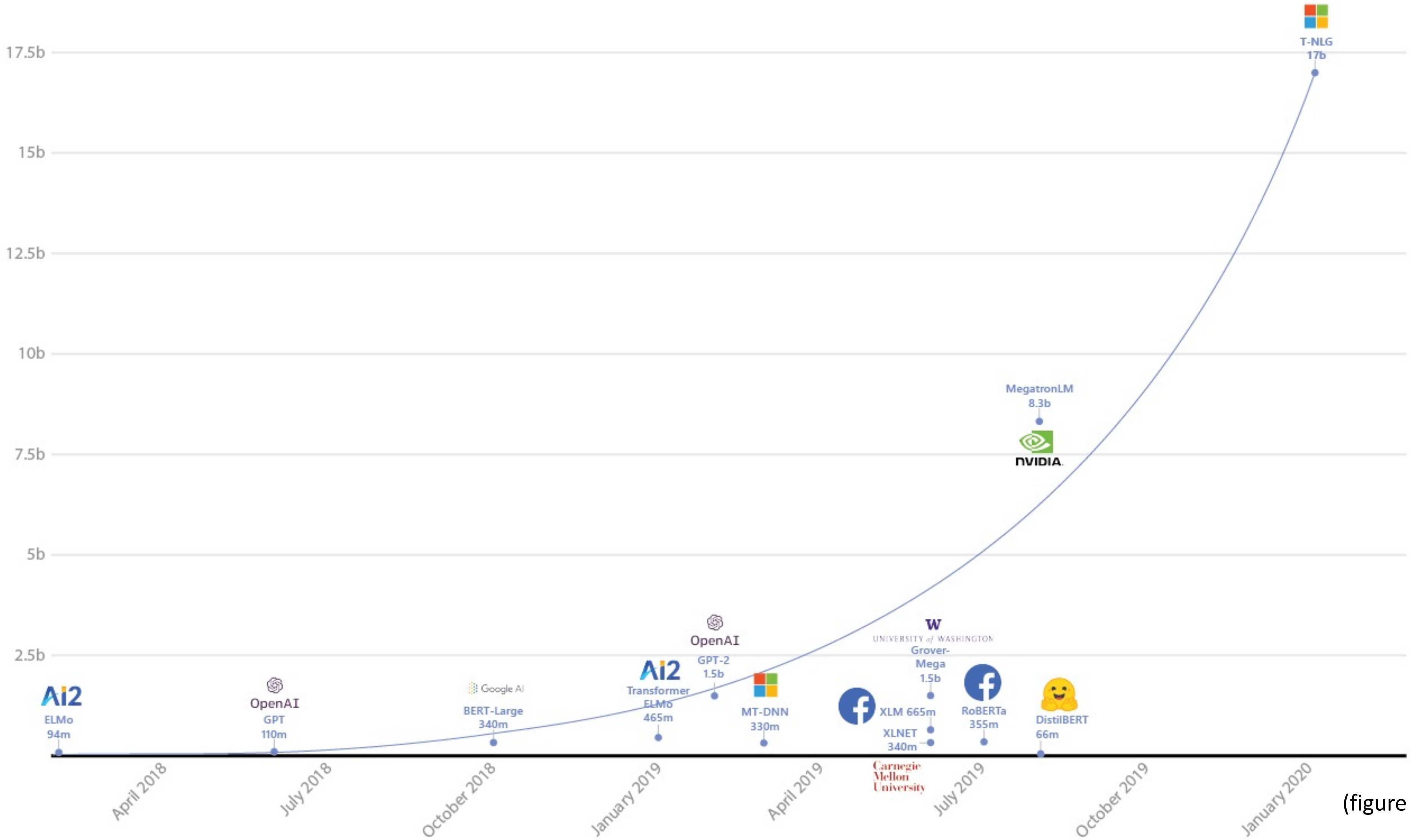
https://gluebenchmark.com/leaderboard

Limitations of Large LMs: Complexity

175b

OpenAI

GPT-3



(figure from Microsoft)

Limitations of Large LMs: Biased Data

Karen was assigned a roommate her first year of college. Her roommate asked her to go to a nearby city for a concert. Karen agreed happily. The show was absolutely exhilarating.

How does the story end?

Karen became good friends with her roommate.

Karen hated her roommate.

Schwartz, R., Sap, M., Konstas, I., Zilles, L., Choi, Y., & Smith, N.A. (2017). The Effect of Different Writing Tasks on Linguistic Style: A Case Study of the ROC Story Cloze Task. In CoNLL 2017.
Mostafazadeh, N., Chambers, N., He, X., Parikh, D., Batra, D., Vanderwende, L., Kohli, P. & Allen, J. (2016). A corpus and cloze evaluation for deeper understanding of commonsense stories. In NAACL 2016.

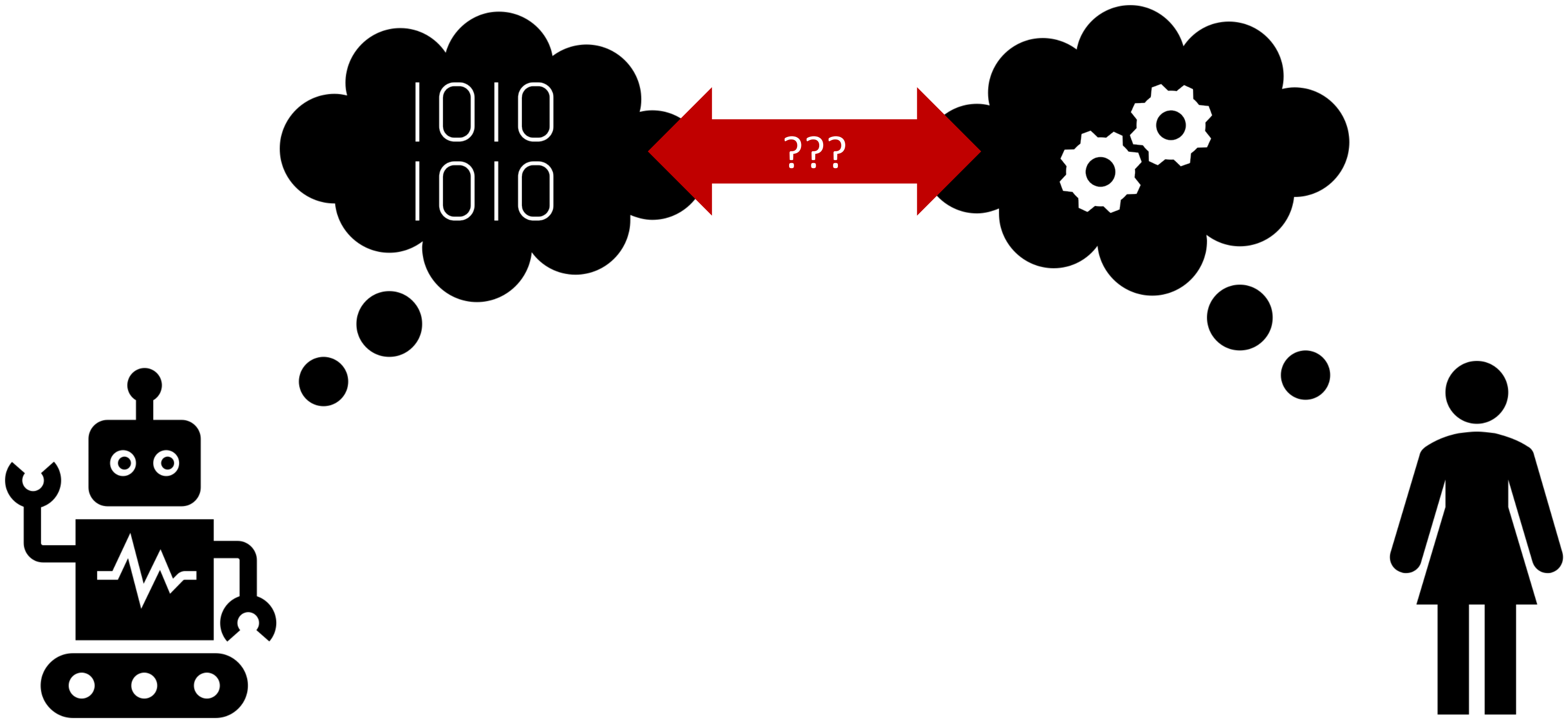
Next Steps

- In order to achieve true commonsense reasoning for natural language understanding (NLU), these key problems will be important to solve:
    1. Better understanding of modeling design choices
    2. External knowledge acquisition and incorporation into system reasoning
    3. Stronger definitions and understanding of system reasoning
    4. Broader, multidimensional metrics for evaluating system reasoning

Storks, S., Gao, Q., & Chai, J.Y. (2020). Recent Advances in Natural Language Inference: A Survey of Benchmarks, Resources, and Approaches. arXiv: 1904.011672 [cs.CL].

Key Questions

1. Is the underlying “reasoning” of large LMs **coherent**?
    - Logical, consistent, and using same supporting evidence as humans to reach a conclusion
2. How can we support more coherent reasoning in large LMs?



Beyond the Tip of the Iceberg:

Assessing Coherence of Text Classifiers

Shane Storks & 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 Short Paper

Textual Entailment

Dialog:

A_1: Yeah, yeah. Is that why you like aerobics classes, because you're not, sort of, someone else is doing the counting for you, so,

B_1: Yeah.

…

B_2: And, someone else is telling me, okay, you know, let's move this way, let's move that way,

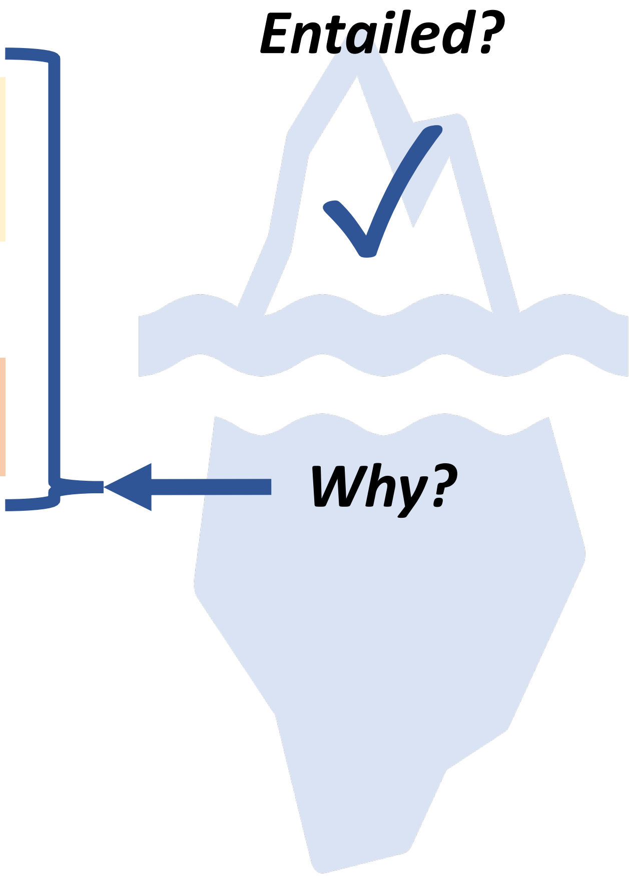
A_2: Uh-huh, uh-huh.

B_3: instead of me having to think about it so much.

…

Hypothesis:

Speaker **B** likes the aspect of Aerobics that someone else is leading.



Zhang, C., & Chai, J.Y. (2010). Towards Conversational Entailment: An Empirical Investigation. In EMNLP 2010.

Coherence

Dialog:

A_1: Well, ironically enough I'm sitting here with a cast on my leg because I resumed an aerobics class the night before last.

B_1: Oh, no.

A_2: I ripped the ligaments in my right ankle.

Hypothesis:

Speaker **A** ripped the ligaments in her ankle at aerobics class.

Accuracy:
full-text correct

Strict Coherence:
all spans correct

Lenient Coherence:
average accuracy on spans

Zhang, C., & Chai, J.Y. (2010). Towards Conversational Entailment: An Empirical Investigation. In EMNLP 2010.

Empirical Results

- Despite high accuracy from SOTA text classifiers, we see significant drops from accuracy to coherence across the board!

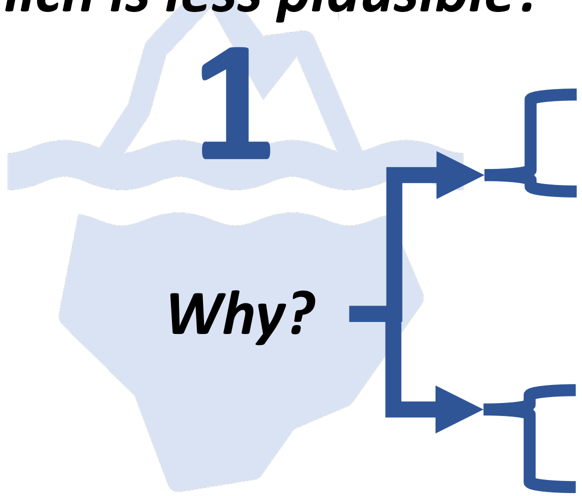
CE, *test*:

Model	**Accuracy** (%)	**Strict Coherence** (Δ; %)	**Lenient Coherence** (Δ; %)
majority	57.8	–	–
BERT	55.8	28.5 (-27.3)	35.7 (-20.1)
RoBERTa	70.9	39.0 (-31.9)	47.5 (-23.4)
↪ + MNLI	78.5	50.6 (-27.9)	58.2 (-20.3)
DeBERTa	67.4	37.2 (-30.2)	45.2 (-22.2)

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
Williams, A., Nangia, N., & Bowman, S.R. (2018). A Broad-Coverage Challenge Corpus for Sentence Understanding through Inference. NAACL HLT 2017.
He, P., Liu, X., Gao, J., & Chen, W. (2021). DeBERTa: Decoding-enhanced BERT with Disentangled Attention. arXiv: 2006.03654.

Abductive Reasoning in narrative Texts (ART)

Which is less plausible?



Story 1:

Kelly wanted to try out for soccer this year.
Kelly tried out for the soccer team but was cut.
Kelly celebrated by getting pizza.

Story 2:

Kelly wanted to try out for soccer this year.
Kelly made it onto the team.
Kelly celebrated by getting pizza.

Bhagavatula, C., Le Bras, R., Malaviya, C., Sakaguchi, K., Holtzman, A., Rashkin, H., Downey, D., Yih, S.W., & Choi, Y. (2020). Abductive commonsense reasoning. In ICLR 2020.

Empirical Results

- Despite high accuracy from SOTA text classifiers, we see significant drops from accuracy to coherence across the board!

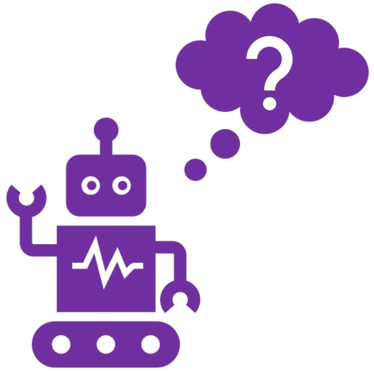
ART, *validation*:

Model	**Accuracy** (%)	**Strict Coherence** (Δ; %)	**Lenient Coherence** (Δ; %)
majority	55.0 (50.1)	–	–
BERT	66.7 (66.7)	42.3 (-24.4)	43.7 (-23.0)
RoBERTa	87.8 (84.2)	55.0 (-32.8)	59.3 (-28.5)
DeBERTa	88.4 (85.7)	59.8 (-28.6)	61.8 (-26.6)

Bhagavatula, C., Le Bras, R., Malaviya, C., Sakaguchi, K., Holtzman, A., Rashkin, H., Downey, D., Yih, S.W., & Choi, Y. (2020). Abductive commonsense reasoning. In ICLR 2020.
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.

Summary

- We proposed a quick, effective, and versatile paradigm for measuring the coherence of a text classifier’s predictions
  - Unlock strong insights from small amount of annotation!
- On selected NLU tasks, SOTA pre-trained LMs perform incoherent reasoning based on spurious intermediate evidence



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)

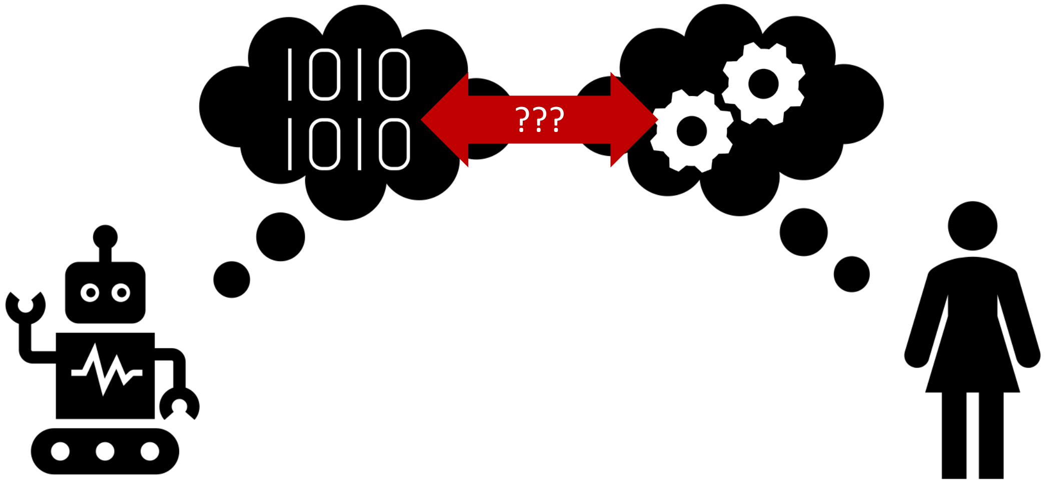
University of Michigan, Computer Science and Engineering Division

sstorks@umich.edu

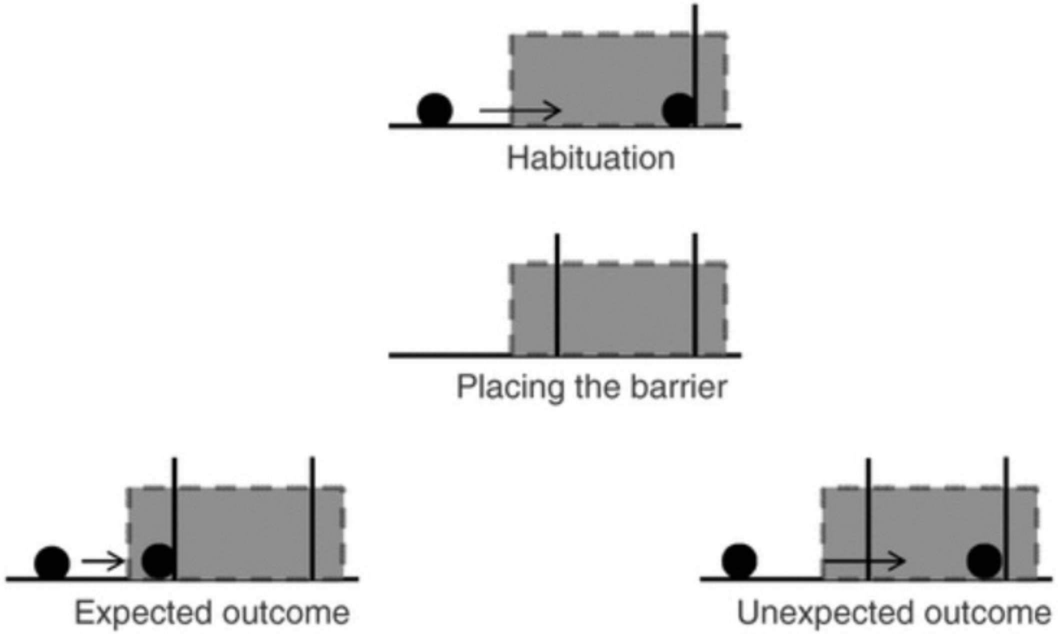
Findings of EMNLP 2021 Long Paper

Motivation

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



Physical Commonsense




(Parents.com)


(dreamstime)

Bliss, J. (2008). Commonsense reasoning about the physical world. In *Studies in Science Education*, 44(2): 123-155.
Lake, B., Ullman, T.D., Tenenbaum, J.B., & Gershman, S.J. (2017). Building machines that learn and think like people. In *Behavioral and Brain Sciences,* 40.
Hespos, S.J. & vanMarle, K. (2011). Physics for infants: characterizing the origins of knowledge about objects, substances, and number.

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. <u>Ann wrote in the book.</u>

Story B

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

!

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)

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

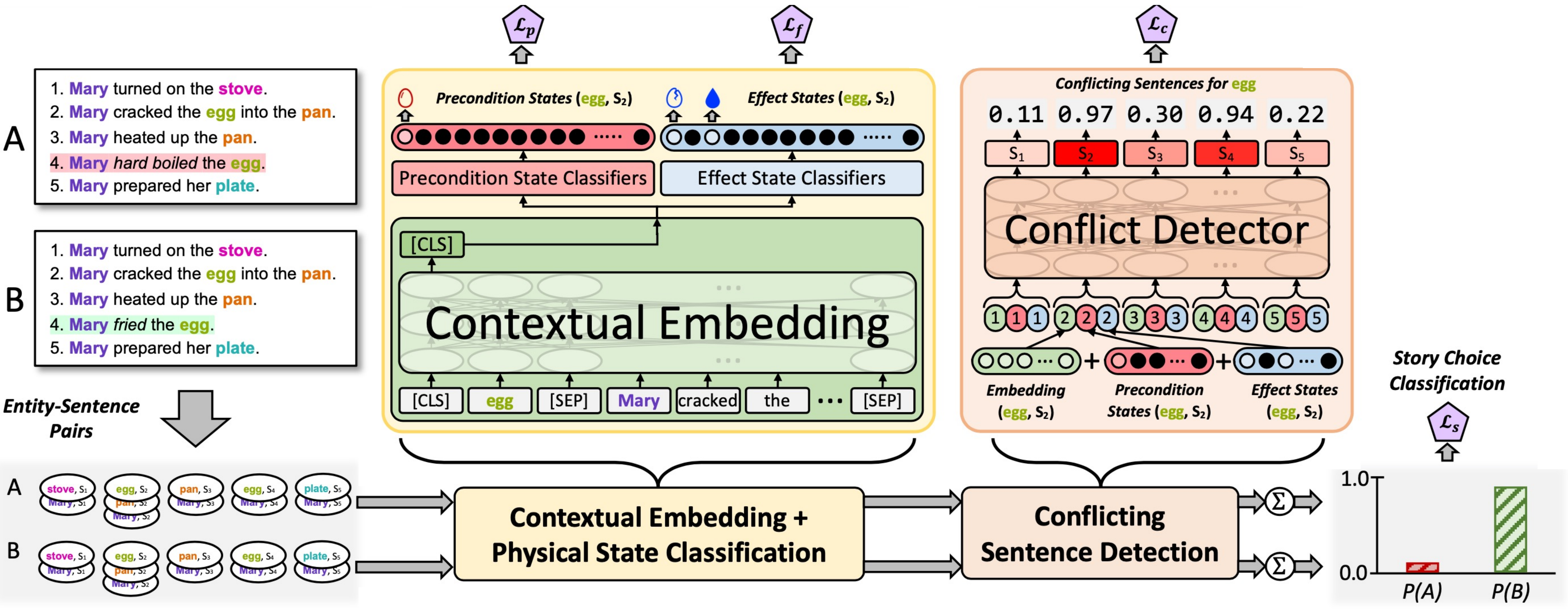
- ***Objects***
  1. Location
  2. Exist
  3. Clean
  4. Power
  5. Functional
  6. Pieces
  7. Wet
  8. Open
  9. Temperature
  10. Solid
  11. Contain
  12. Running
  13. Moveable
  14. Mixed
  15. Edible

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

Evaluation Metrics

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

Tiered Baseline



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

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.

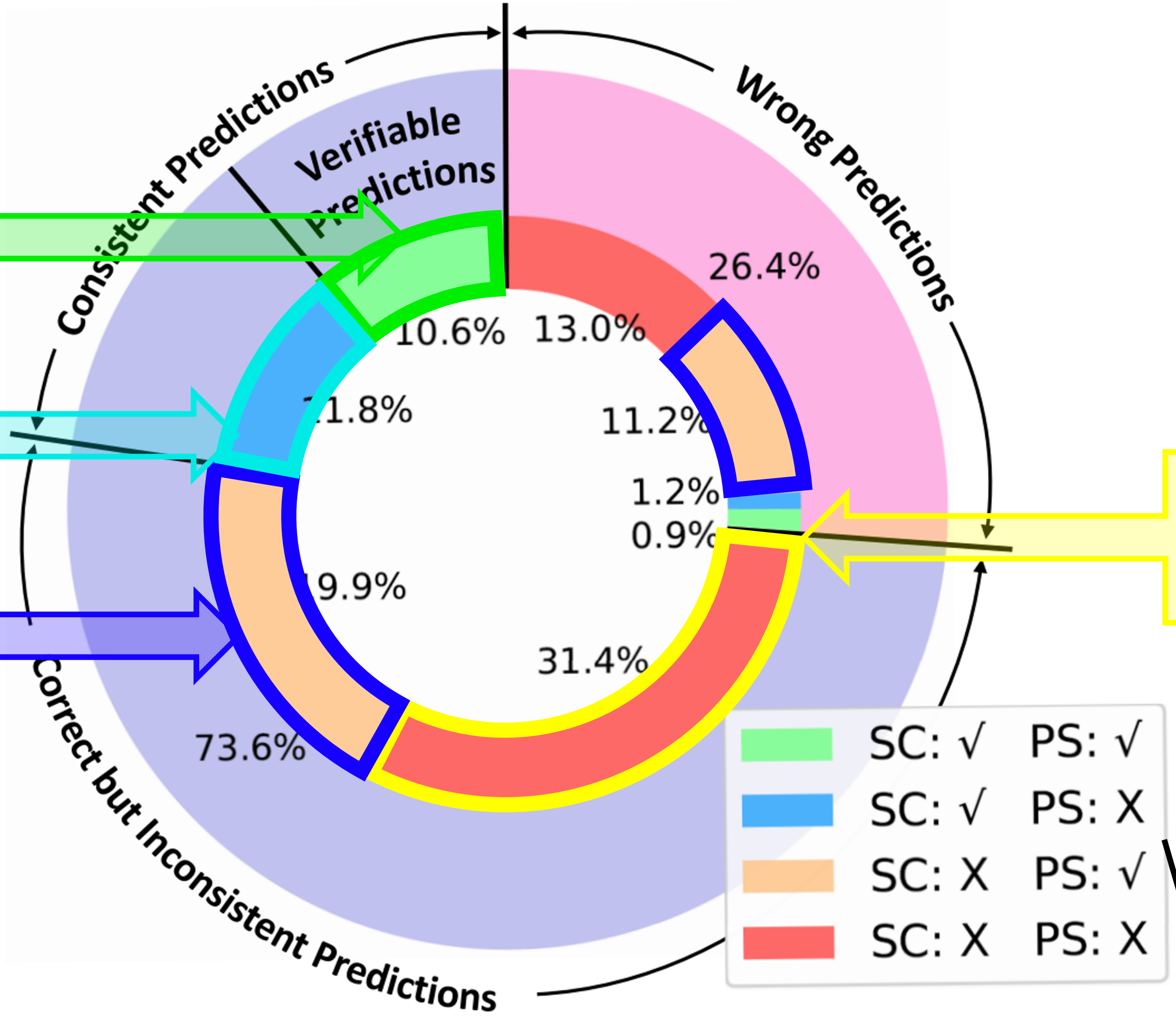
He, P., Liu, X., Gao, J., & Chen, W. (2021). DeBERTa: Decoding-enhanced BERT with Disentangled Attention. arXiv: 2006.03654.

Error Distribution

Correct and entirely verifiable!

Consistent but not verifiable!

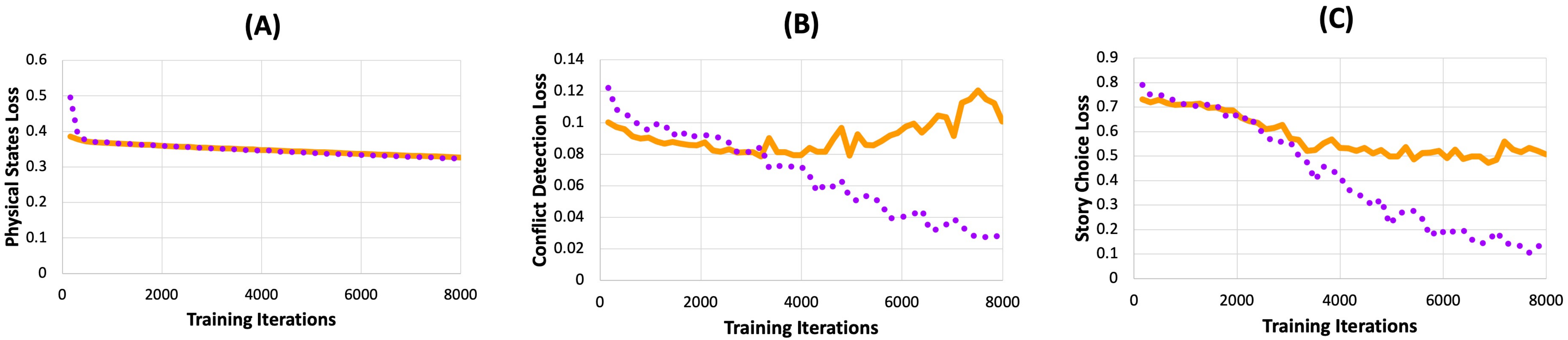
Correct states, but unsuccessful conflict detection. 🤔



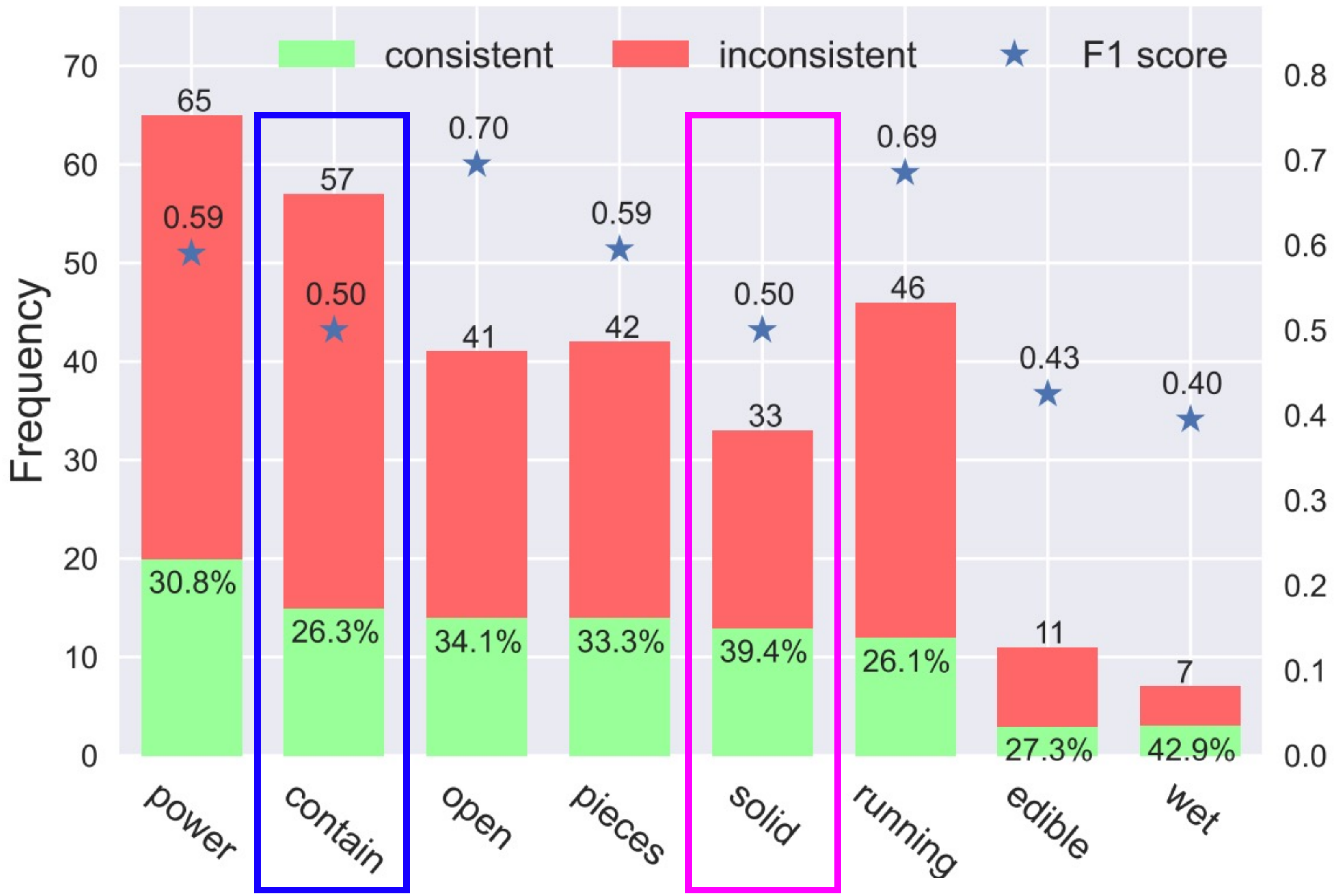
Correct, but entirely unverifiable!

SC: sentence conflict
PS: physical states

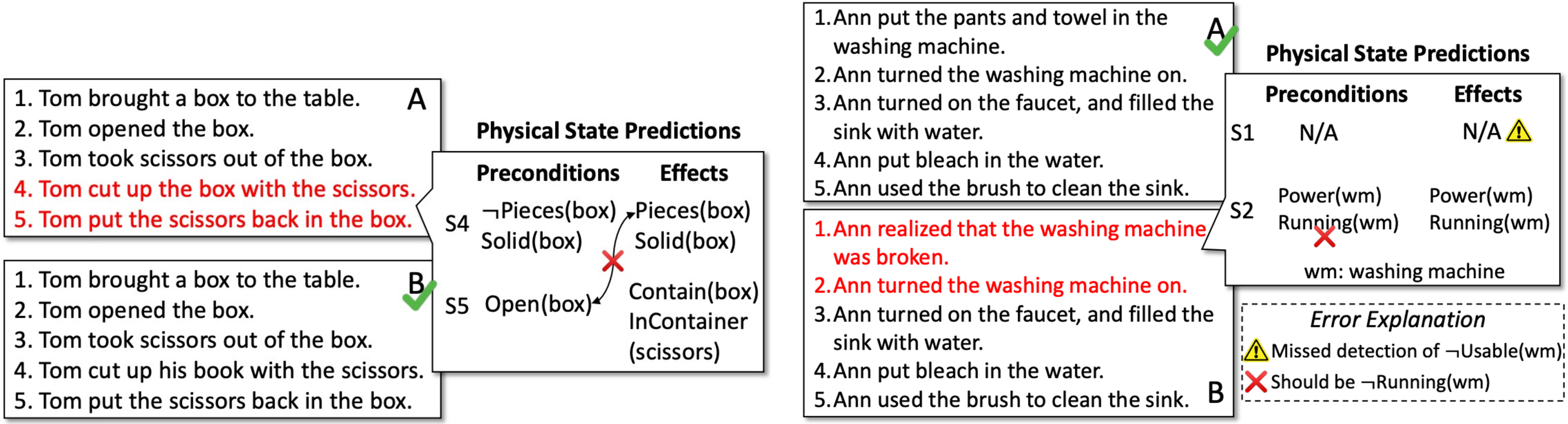
Tiered Task Learning



Utility of Attributes



Sample System Outputs



(a) A verifiable prediction.

(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.

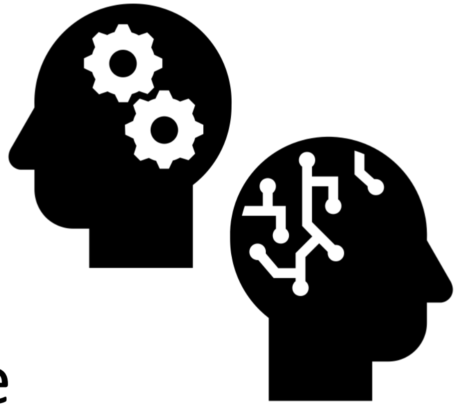
Summary

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.

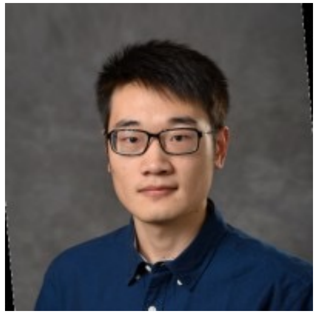
Key Takeaways

1. SOTA systems that perform well on NLU tasks may use incoherent reasoning based on spurious evidence
2. SOTA systems struggle to learn how to reason coherently
   - TRIP provides strong insights for future development of NLU systems with verifiable (physical) commonsense reasoning!
3. Despite exciting SOTA results, incorporating commonsense reasoning into NLU is still a difficult problem ☹



Acknowledgements

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



Thank you!