Using Natural Language Explanations to Incorporate Commonsense Reasoning in Neural Networks

Nazneen Rajani, Senior Research Scientist, Salesforce

nazneen.rajani@salesforce.com, @NazneenRajani

March 3, 2020

Geoffrey Hinton
@geoffreyhinton

Suppose you have cancer and you have to choose between a black box AI surgeon that cannot explain how it works but has a 90% cure rate and a human surgeon with an 80% cure rate. Do you want the AI surgeon to be illegal?

12:37 PM · Feb 20, 2020 · Twitter Web App

Outline

Explain Yourself! Leveraging Language Models for Commonsense Reasoning

Nazneen Fatema Rajani **Bryan McCann** **Caiming Xiong** **Richard Socher**
Salesforce Research
Palo Alto, CA, 94301
{nazneen.rajani,bmccann,cxiong,rsocher}@salesforce.com

ACL 2019

ESPRIT: Explaining Solutions to Physical Reasoning Tasks

Anonymous ACL submission

Under review at ACL 2020

Commonsense Reasoning

"It will fall down"

Explanation: "gravity"

Commonsense Reasoning in Neural Networks

- Neural networks lack commonsense reasoning abilities.
(Talmor et al., 2019; Zellers et al., 2019, Bisk et al., 2019)
- Approximately 30 accuracy points behind humans.

Commonsense Reasoning Tasks

The person blows the leaves from a grass area using the blower. The blower...

- a) puts the trimming product over her face in another section.
- b) is seen up close with different attachments and settings featured.
- c) continues to blow mulch all over the yard several times.
- d) blows beside them on the grass.

SWAG (Zellers et al., 2018)

[Goal] Make an outdoor pillow
[Sol1] Blow into a tin can and tie with rubber band ✘
[Sol2] Blow into a trash bag and tie with rubber band ✔

[Goal] To make a hard shelled taco,
[Sol1] put seasoned beef, cheese, and lettuce onto the hard shell. ✘
[Sol2] put seasoned beef, cheese, and lettuce into the hard shell. ✔

[Goal] How do I find something I lost on the carpet?
[Sol1] Put a solid seal on the end of your vacuum and turn it on. ✘
[Sol2] Put a hair net on the end of your vacuum and turn it on. ✔

PIQA (Bisk et al., 2019)

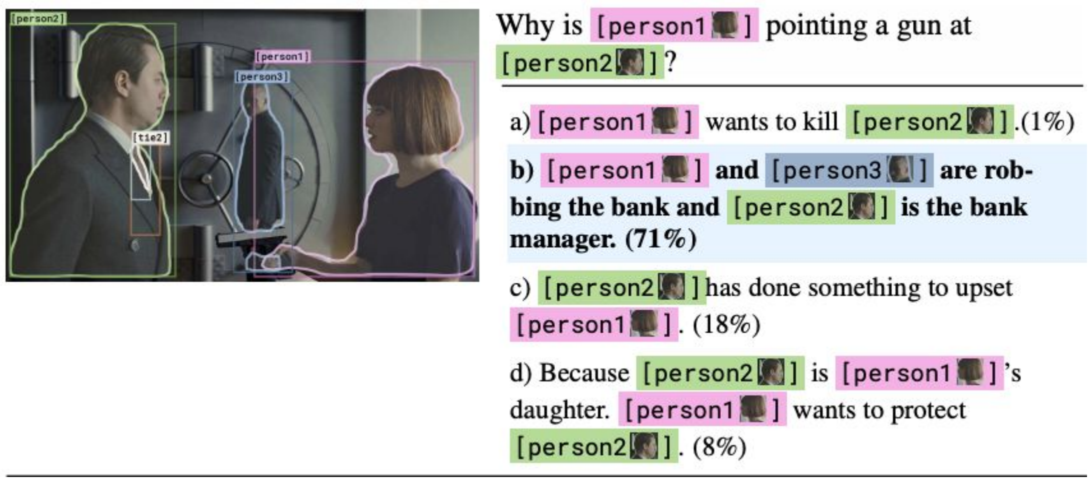
Where would I not want a fox?
👍 hen house, 👎 england, 👎 mountains, 👎 english hunt, 👎 california

Commonsense QA (Talmor et al., 2019)

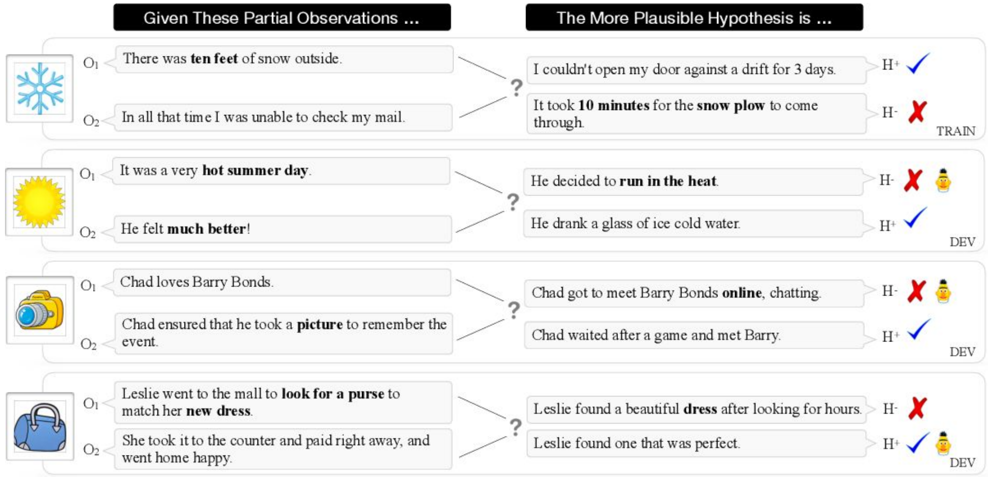
Context	Right Ending	Wrong Ending
Karen was assigned a roommate her first year of college. Her roommate asked her to go to a nearby city for a concert. Karena greed happily. The show was absolutely exhilarating.	Karen became good friends with her roommate	Karen hated her roommate

Story Cloze (Mostafazadeh et al., 2016)

And more...



Visual Commonsense Reasoning
(Zellers et al., 2019)



Abductive Commonsense Reasoning
(Bhagavatula et al., ICLR 2020)

Problem Statement

How can we incorporate commonsense reasoning in Neural Networks?

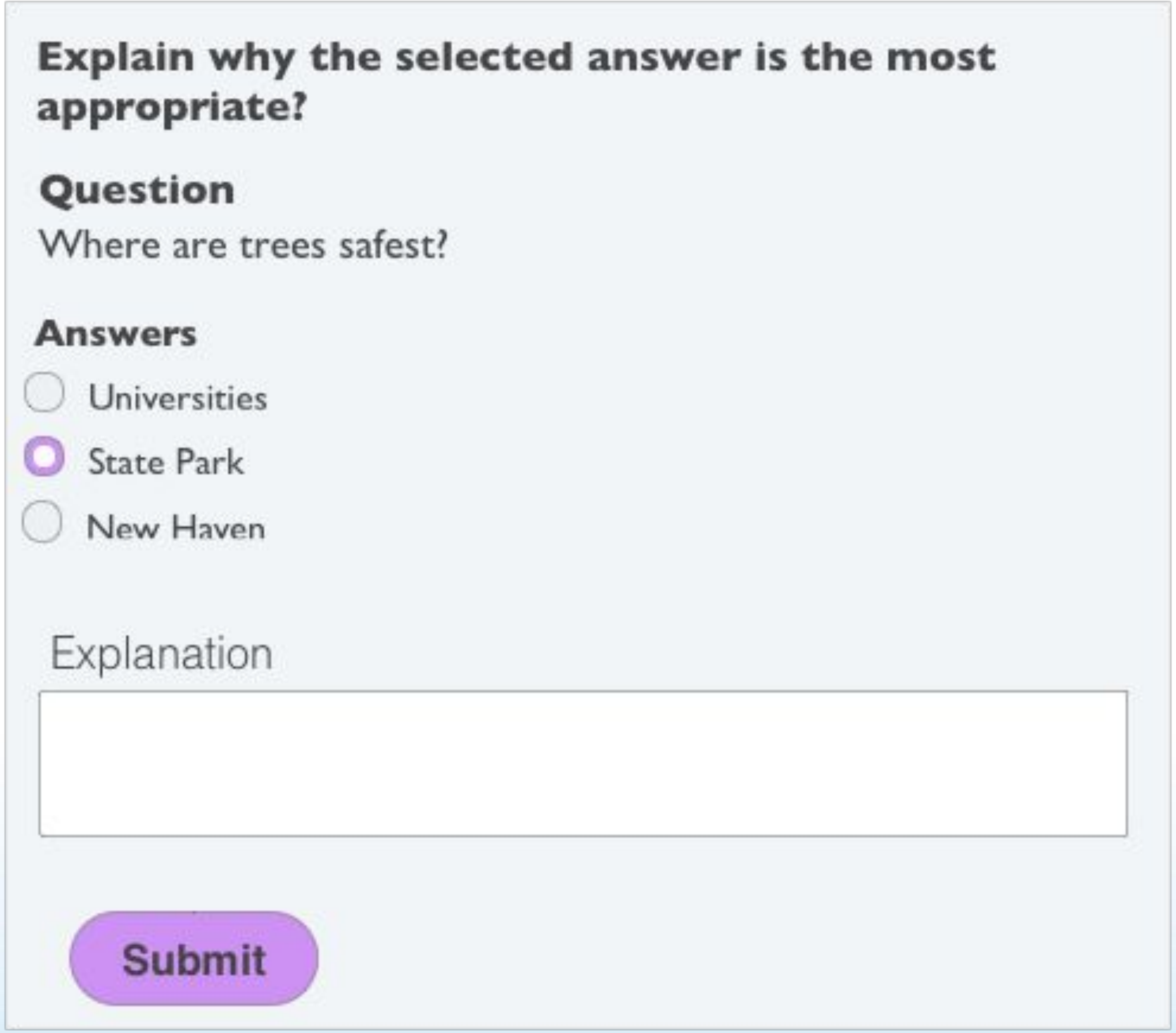
Question | **Can neural networks use human commonsense explanations?**

Question | Can neural networks generate their own commonsense explanations?

Question | Can neural networks use their own auto-generated explanations?

Question | Can neural commonsense explanations transfer between tasks?

Human Commonsense Explanations


Explain why the selected answer is the most appropriate?

Question

Where are trees safest?

Answers

- Universities
- State Park
- New Haven

Explanation

Submit

Human Commonsense Explanations

Explain why the selected answer is the most appropriate?

Question

Where are trees safest?

Answers

- Universities
- State Park
- New Haven

Explanation

State Parks have rules to keep trees protected.

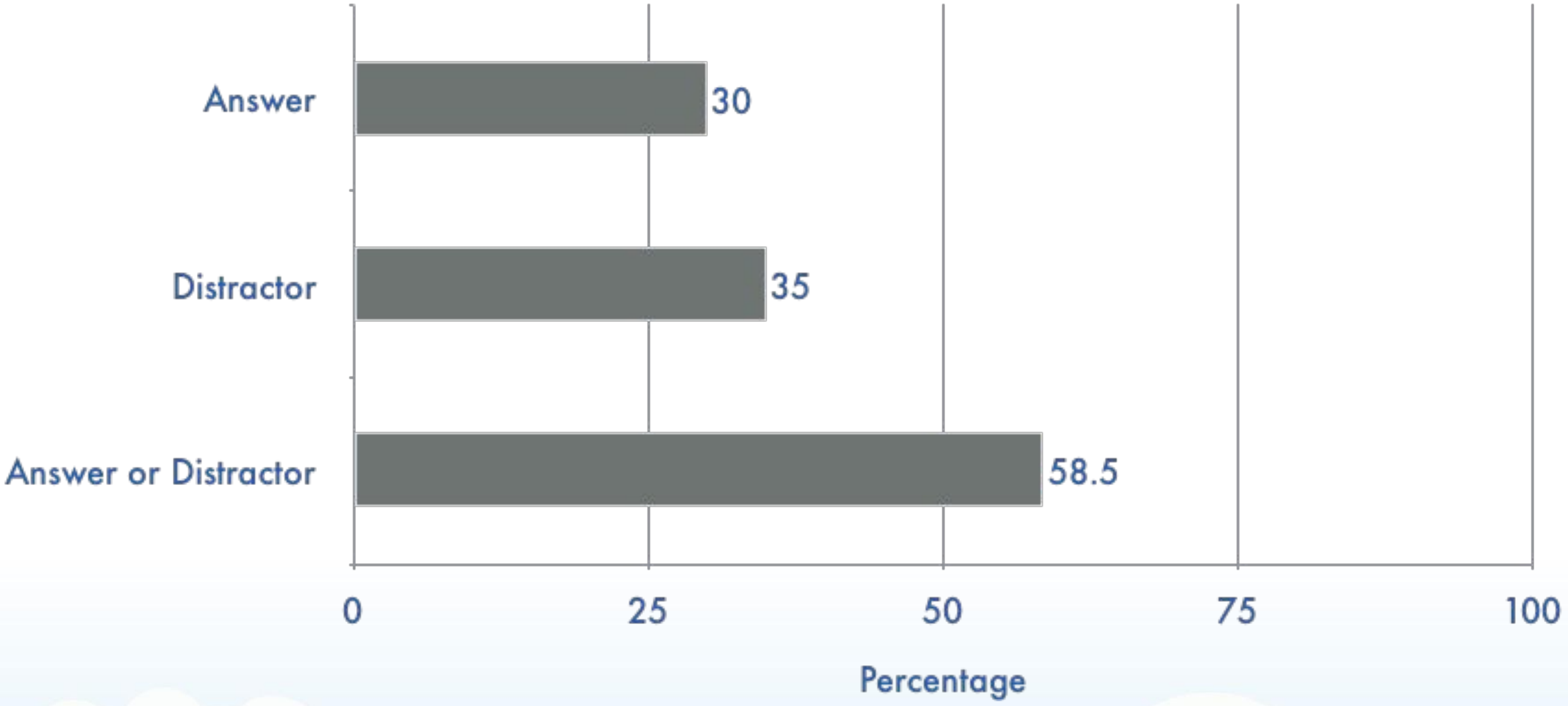
Submit

Human Explanations

- Captured World Knowledge.

Type	Examples
Cause and effects	“disagreements lead to fights”
Social norms	“forgiving activates good karma”
Laws of Physics	“gravity makes things fall”
Geography	“Minnesota is the only option that is a state”
Other	“National parks have rules to protect trees”

Human Explanations Analysis



Can NNs use human commonsense explanations?

salesforce

- Human explanations along with Q and A choices.
- Used by classifier only during training (7610 examples).
- Improved SOTA by 6% points.
- Publicly available:

https://github.com/salesforce/cos-e

Question | Can neural networks use human commonsense explanations?

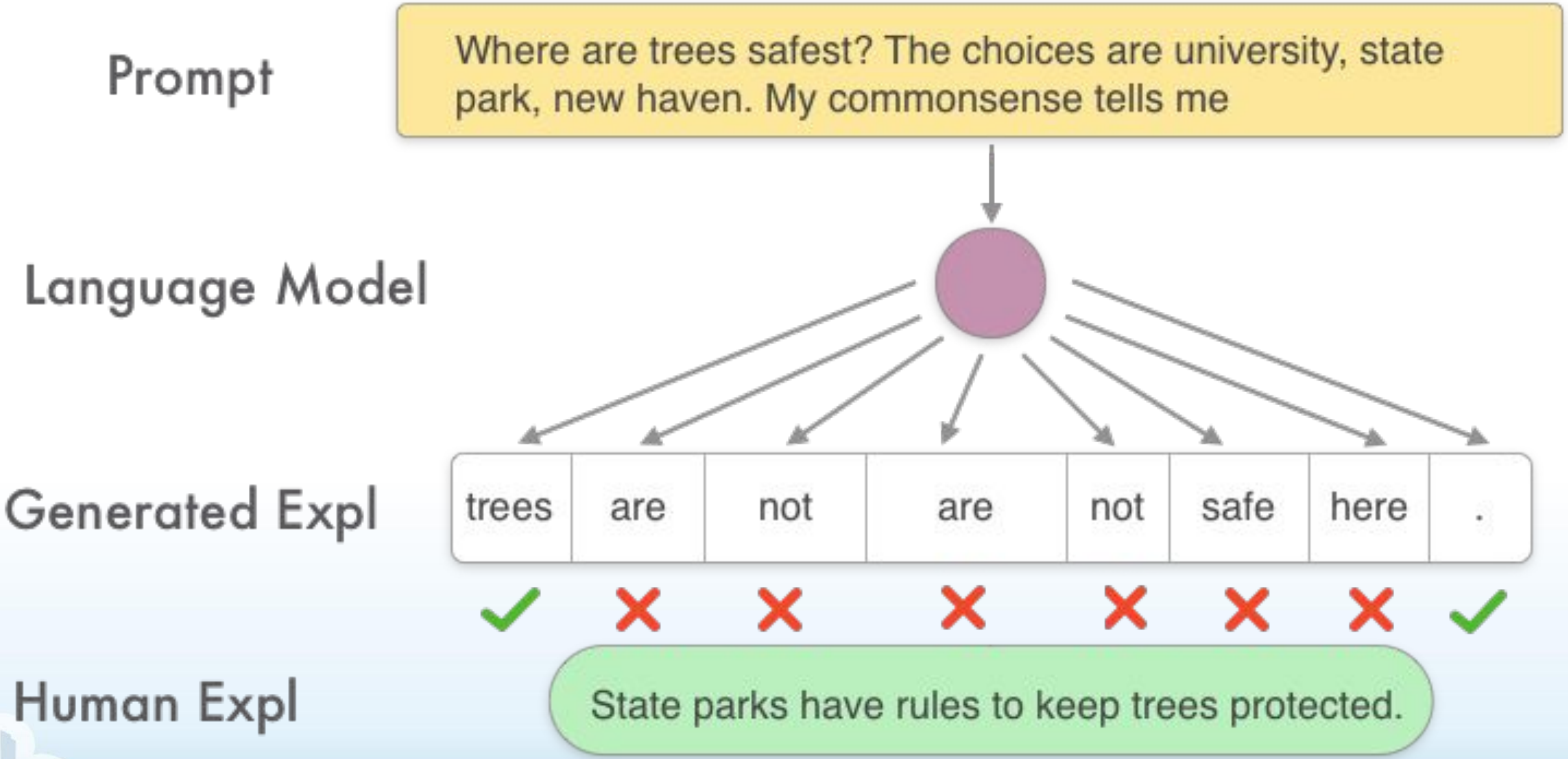
Question | **Can neural networks generate their own commonsense explanations?**

Question | Can neural networks use their own auto-generated explanations?

Question | Can neural commonsense explanations transfer between tasks?

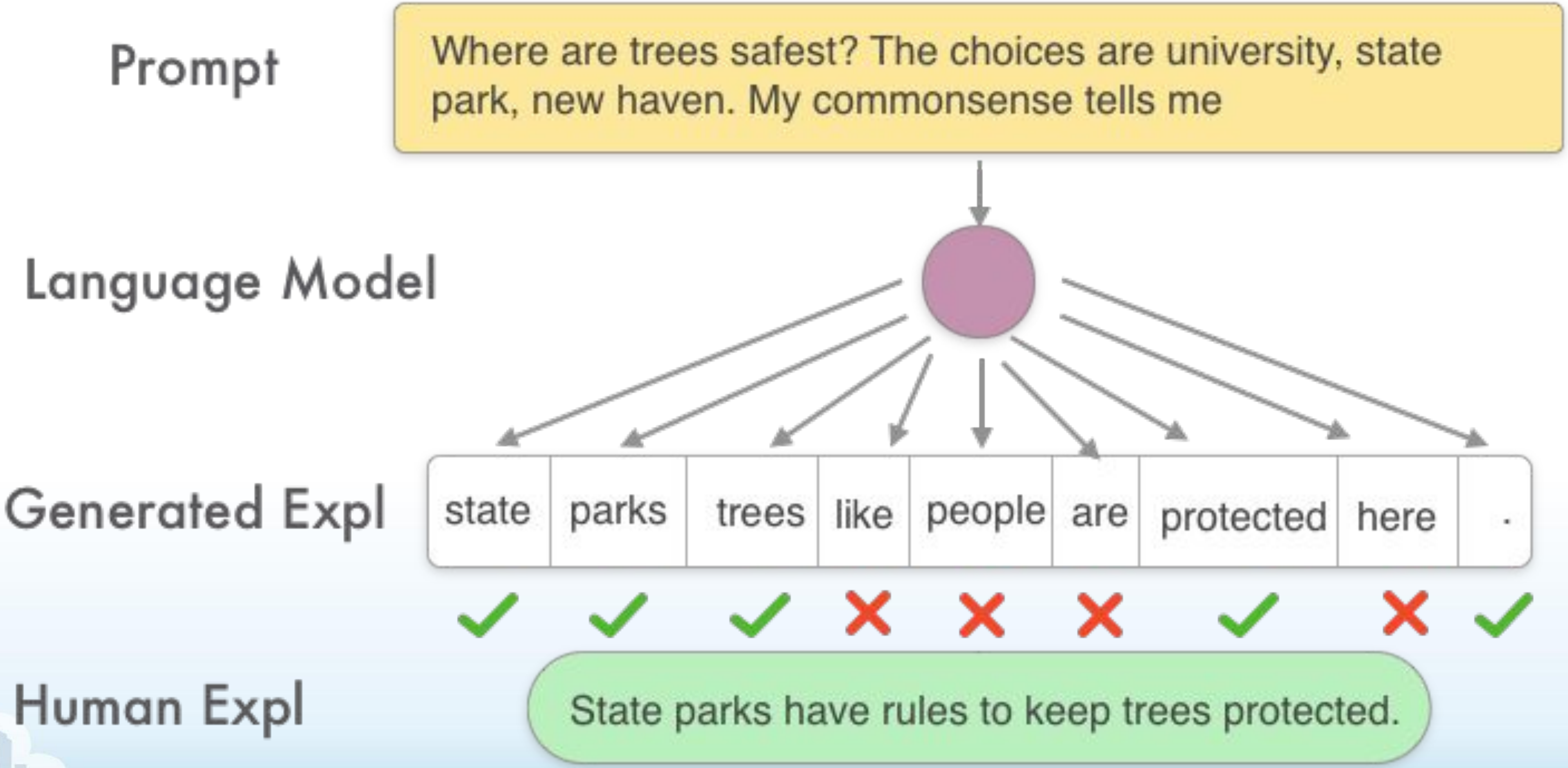
Explanations Generation Model

- Fine-tune language model (LM) on small number of human explanations.



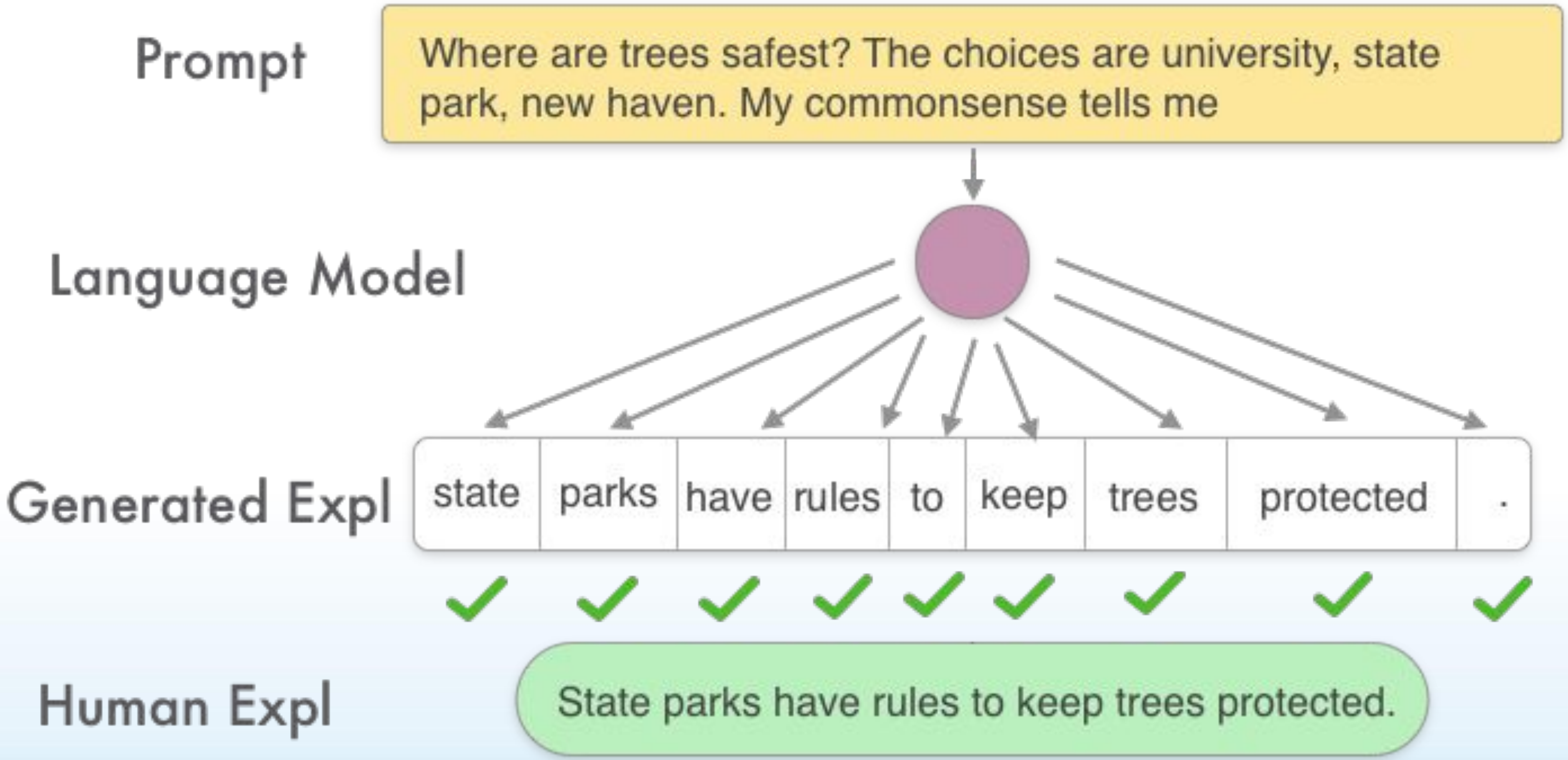
Explanations Generation Model

- Fine-tune language model (LM) on small number of human explanations.



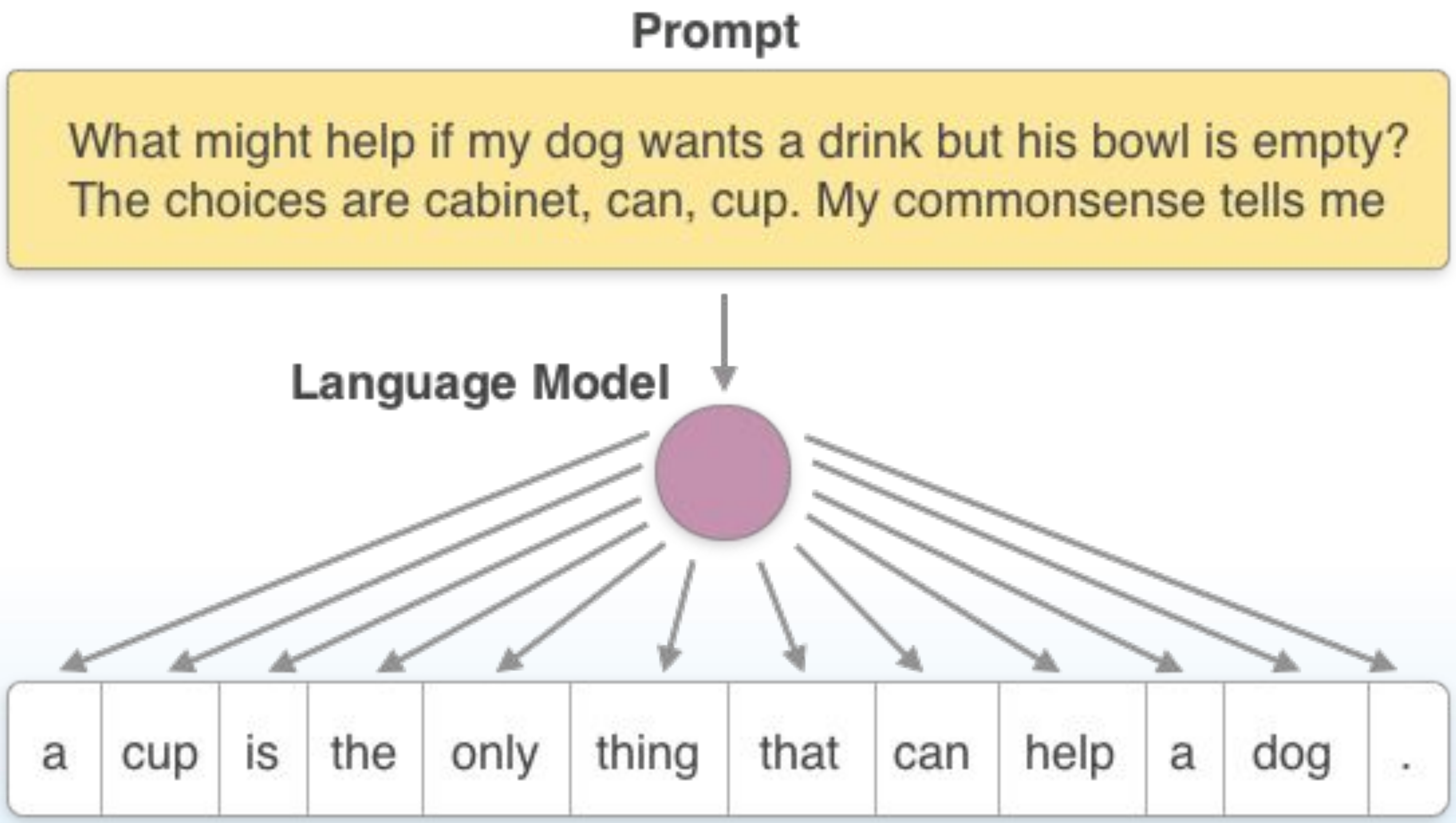
Explanations Generation Model

- Fine-tune language model (LM) on small number of human explanations.



Can NNs generate their own explanation?

- Use fine-tuned LM to generate expl on new instances.



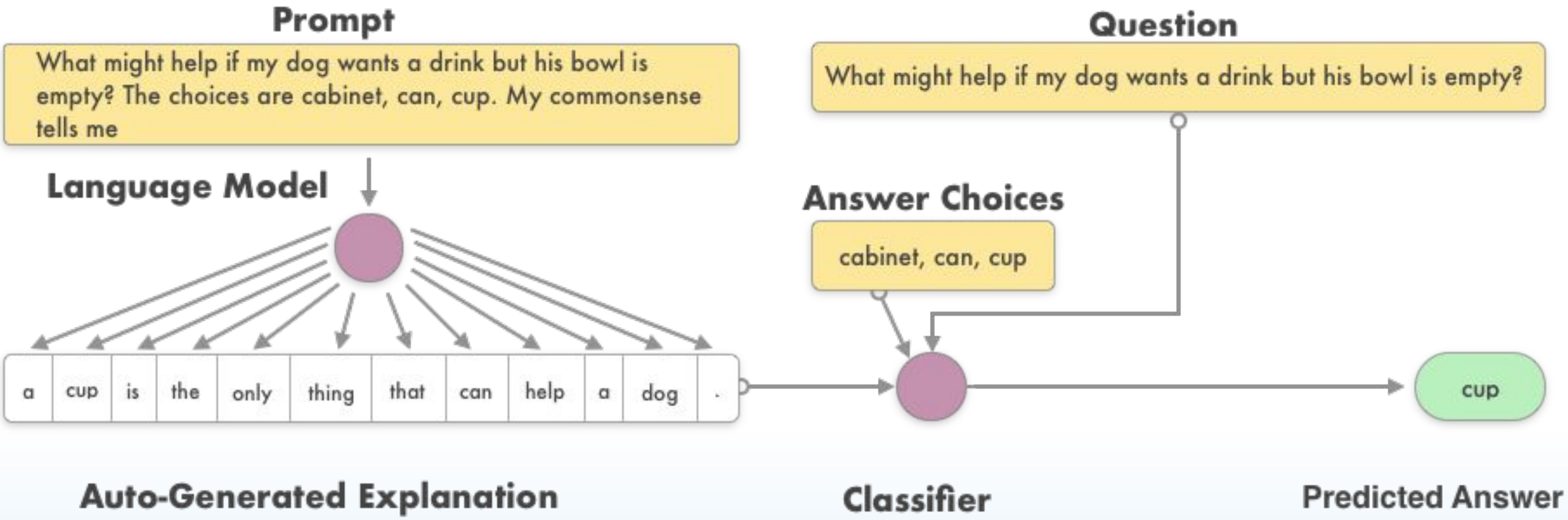
Question | Can neural networks use human commonsense explanations?

Question | Can neural networks generate their own commonsense explanations?

Question | **Can neural networks use their own auto-generated explanations?**

Question | Can neural commonsense explanations transfer between tasks?

Can NNs use their own explanation?



Can NNs use their own explanation?

- Pre-trained GPT as Language Model for explanations
- LM prompt:

```
Q, c0, c1, or c2? My commonsense tells me
```

- BERT for classification:

`[CLS]` Question `[SEP]` Expl `[SEP]` Choice 0 `[SEP]`

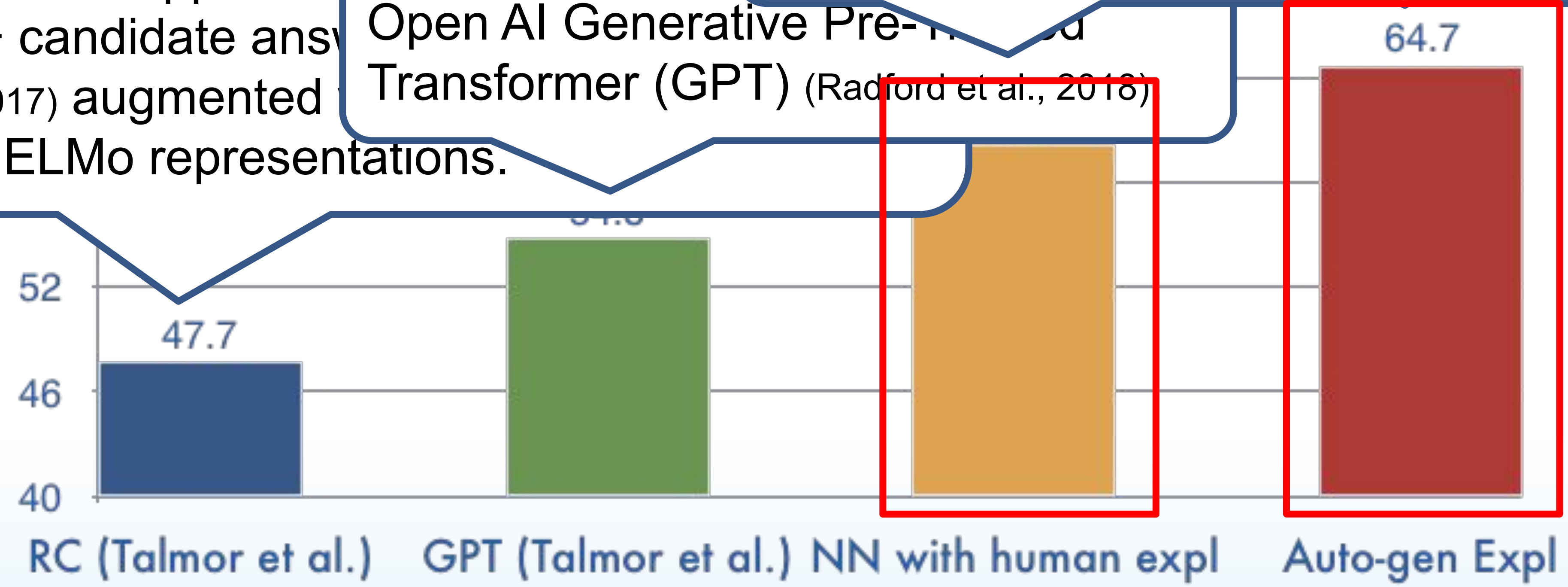
Results

Auto-generated LM explanations during training and inference.

Human commonsen
used during training

Top 100 snippets fr
“Q + candidate ansv
al., 2017) augmented
and ELMo representations.

Open AI Generative Pre-
Transformer (GPT) (Radford et al., 2018)



Error Analysis

Question:	What is the main purpose of having a bath?
Choices:	**cleanness**, use water, exfoliation, <u>hygiene</u>, wetness
Explanation:	the only purpose of having a bath is to **clean** yourself.
Question:	Where can you store your spare linens near your socks?
Choices:	cabinet, <u>chest,</u> hospital, **dresser drawers,** home
Explanation:	**dresser drawers i**s the only place that you can store linens
Question:	Where do you find the most amount of leafs?
Choices:	**forest**, floral arrangement, <u>compost pile</u>, field, ground
Explanation:	The most likely place to find leafs is in a **garden**

Question | Can neural networks use human comonsense explanations?

Question | Can neural networks generate their own commonsense explanations?

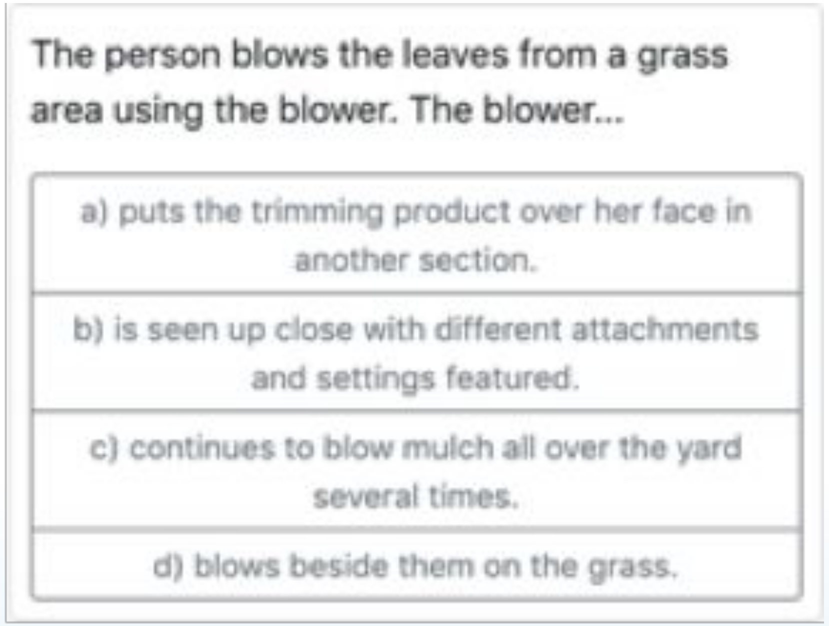
Question | Can neural networks use their own auto-generated explanations?

Question | **Can neural commonsense explanations transfer between tasks?**

Can neural commonsense explanations transfer between tasks?

- LM fine-tuned on CQA explanations.
- Generate explanations on new task domain.

Commonsense QA (Talmor et al., 2019)



SWAG (Zellers et al., 2018)

Can neural commonsense explanations transfer between tasks?

Method	SWAG
BERT	84.2
+ expl transfer	83.6

Question The man examines the instrument in his hand.

Choices The person studies a picture of the man playing the violin.,
The person holds up the violin to his chin and gets ready.,
The person stops to speak to the camera again.,
The person puts his arm around the man and backs away.

Explanation the person is holding the instrument in his hand.

Takeaways

- Human expl used only during training improves performance.
- Expl are a way to incorporate commonsense in NNs.
- LMs are powerful enough to generate meaningful commonsense expl.
- Auto-generated expl improve accuracy by 10% points on CQA.

Neural Networks do not have a causal coherent understanding of the real world

Explaining Solutions to Physical Reasoning Tasks

Under review at ACL 2020

In collaboration with:

Rui Zhang (r.zhang@yale.edu)
Yi Chern Tan (yichern.tan@yale.edu)
Jeremy Weiss (j.weiss@yale.edu)
Aadit Vyas (aadit.vyas@yale.edu)
Abhijit Gupta (abhijit.gupta@yale.edu)
Dragomir Radev (dragomir.radev@yale.edu)

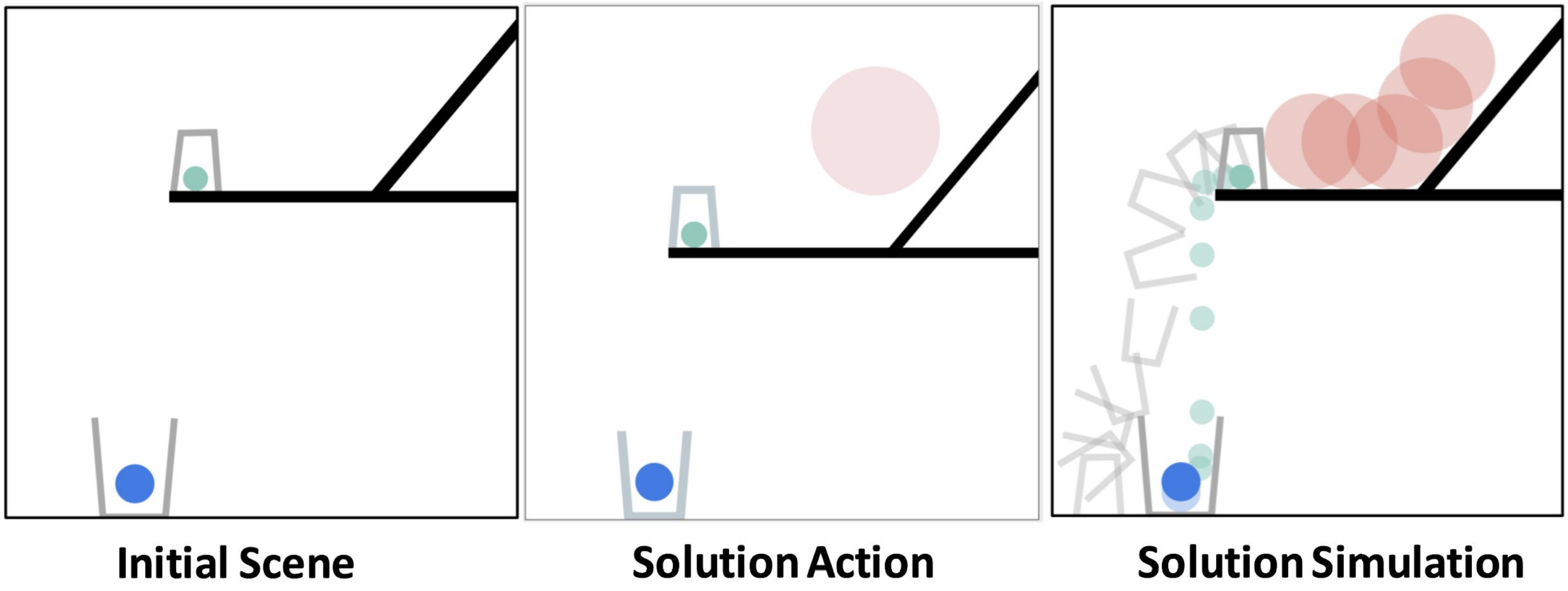
Introduction

- Humans can reason about qualitative physics but AI systems can't --
  - a falling ball will bounce
  - predict projection of ball and catch it
- **Intuition:** Low dim proxy for the world model focusing on physical concepts.
- **Goal:** Use natural language to explain qualitative physics involved in the AI system's behavior and prediction.
- Three physical concepts:
  - Gravity
  - Friction
  - Collision

Physical Reasoning (PHYRE) tasks

Facebook AI Research (Bakhtin et al., NeurIPS 2019)

Goal: Make the green ball touch the blue ball.

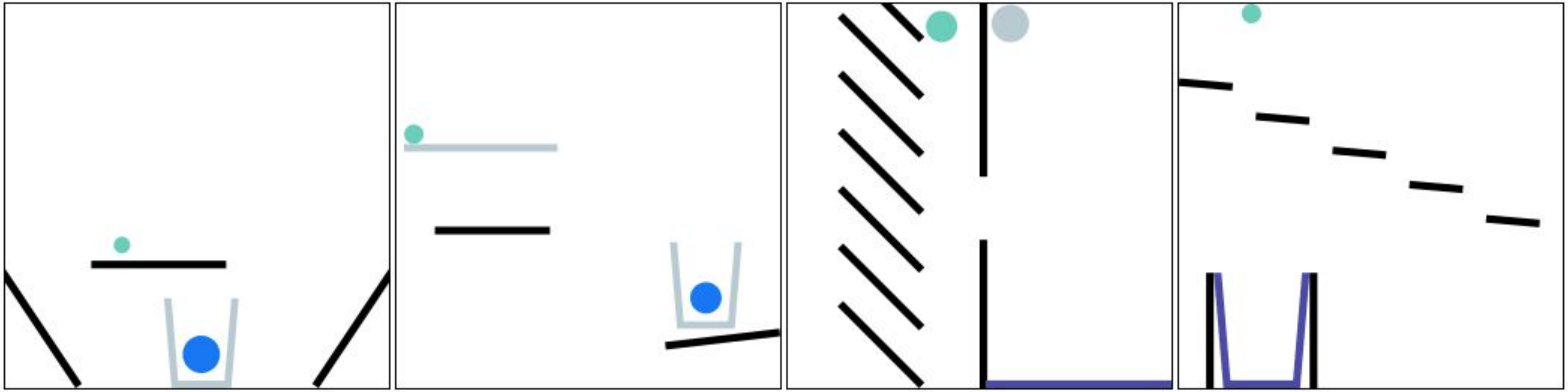


https://github.com/facebookresearch/phyre

PHYRE Benchmark Dataset

- Two tiers (all continuous) with 25 templates each:
  - ✓ Those that require one ball to reach the goal state
  - Those that require two balls to reach the goal state
- Tasks within a template (100 each) have the **same** goal but different initial state.

Make the green ball touch the blue/purple object by adding red objects



https://github.com/facebookresearch/phyre

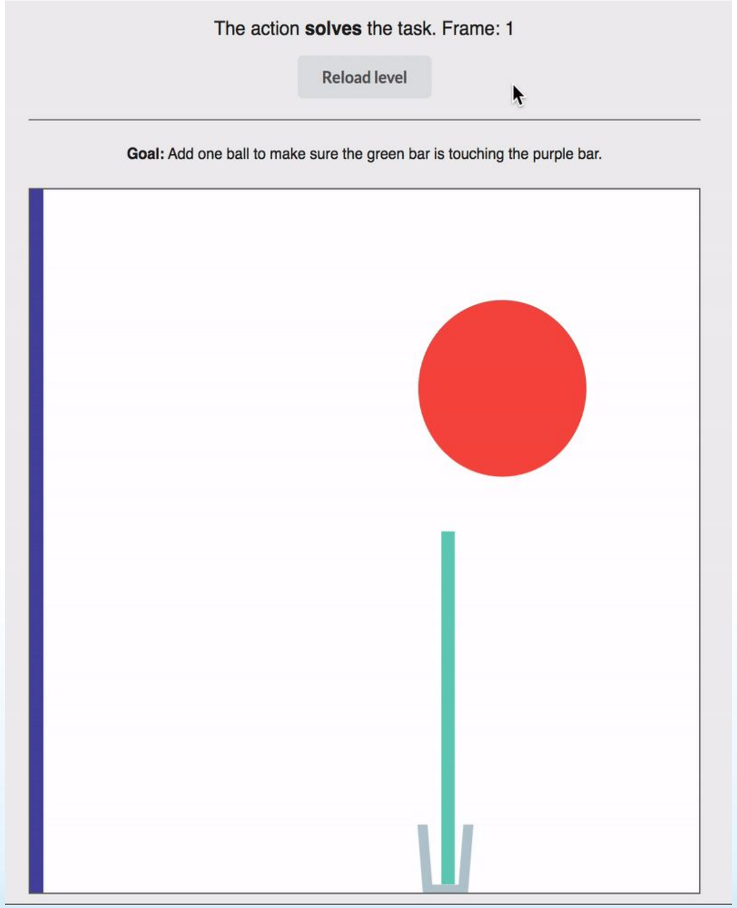
Dataset Annotation

- Manually describing some templates and tasks
  - General enough so that it is useful
- Manual annotation observations:
  - Collision is the most frequent (avg=54) and crucial concept to reach goal state.
  - Initial config is crucial w.r.t. object positions and attributes.
- Two stage process:
  - **Phase 1** - Salient collision detection
  - **Phase 2** - Natural language explanation
    - Initial state description
    - Collision description
- Random split 785 tasks into train/dev/test - 625/84/76

Phase 1: Salient Collision Annotations

- Use phyre simulator to extract objects (relations, positions and attributes) and collisions as a table.
- Frames that are causally related to the placement of the red ball

Frame	Event
1	Start
18	Red ball touches green ball
32	Green bar touches gray jar
50	Red ball touches gray jar
52	Red ball touches floor
66	Green bar touches gray jar
69	Green bar touches purple bar
75	Red ball touches right wall
120	Green bar touches floor



Phase 1: Salient event detection

- Use Mturk to annotate salient events
  - Show goal, initial state, and all collisions from a task
  - “Select all frames that are causally related to the placement of the red ball and necessary to complete the goal”
  - Average = 4 salient collisions selected
- Train a binary classifier to detect salient frames:
  - 13 Features extracted from table related to position, attributes (velocity, dynamic/static).
  - Training data 4,851 collisions (793 positive, 4,058 negative).
  - Test data 4,428 collisions (737 positive, 3691 negative).

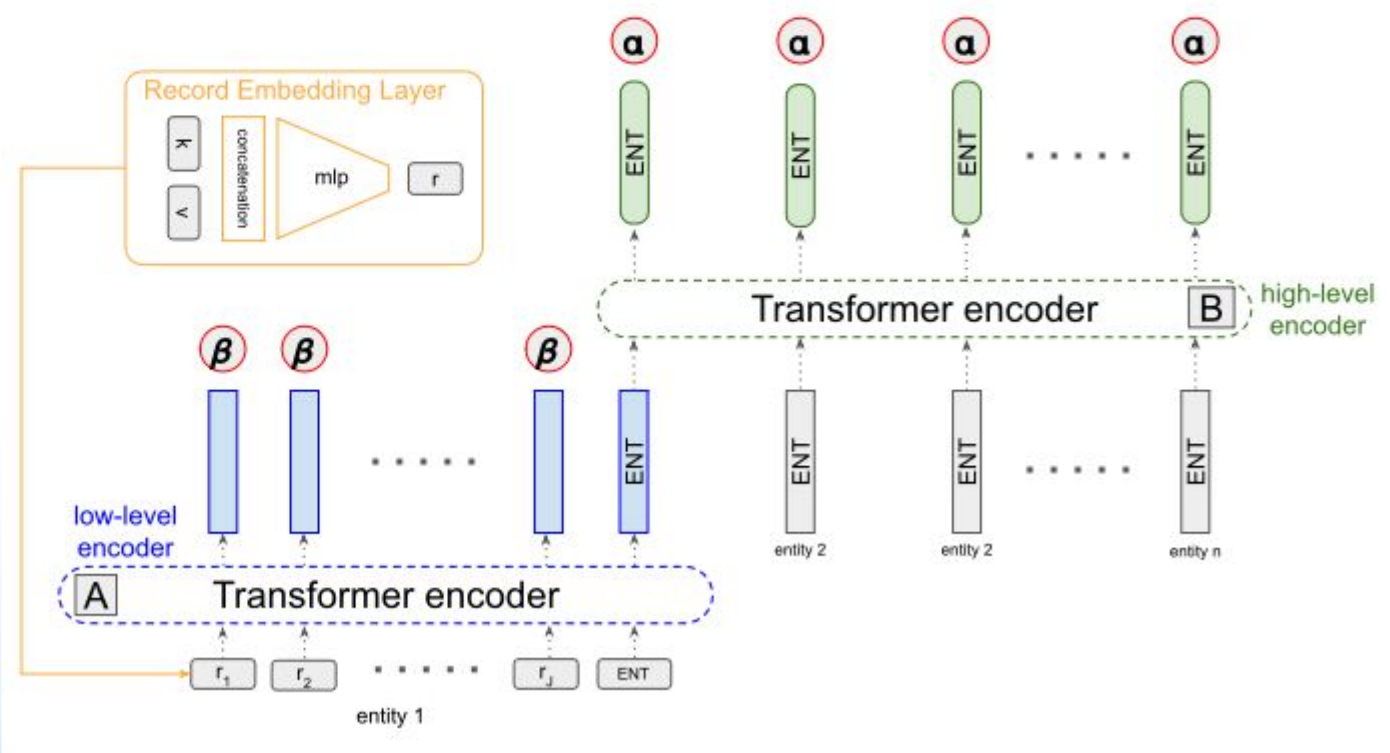
	Precision	Recall	F1
Positive	0.17	1.00	0.29
Decision Tree	0.99	0.96	0.97
Support Vector	0.99	0.97	0.98

Phase 2: Salient frame descriptions

- Use Mturk to collect open-ended description of:
  - **Step 1:** Initial frame
  - **Step 2:** Salient frames detected in phase 1
  - They are shown goal, initial state for step 1 and also the salient frames for step 2.
  - Average 40 words with vocab size of 867.
- Structured data (table) to text generation problem.
- LM generation problem.

Structured data-to-text Model

- Encoder:
  - Learn a record embedding
  - Output:
    - **Avg** of record embeddings
    - **BiLSTM** concat of record embeddings
- Decoder:
  - Hierarchical attention - first entities then their records

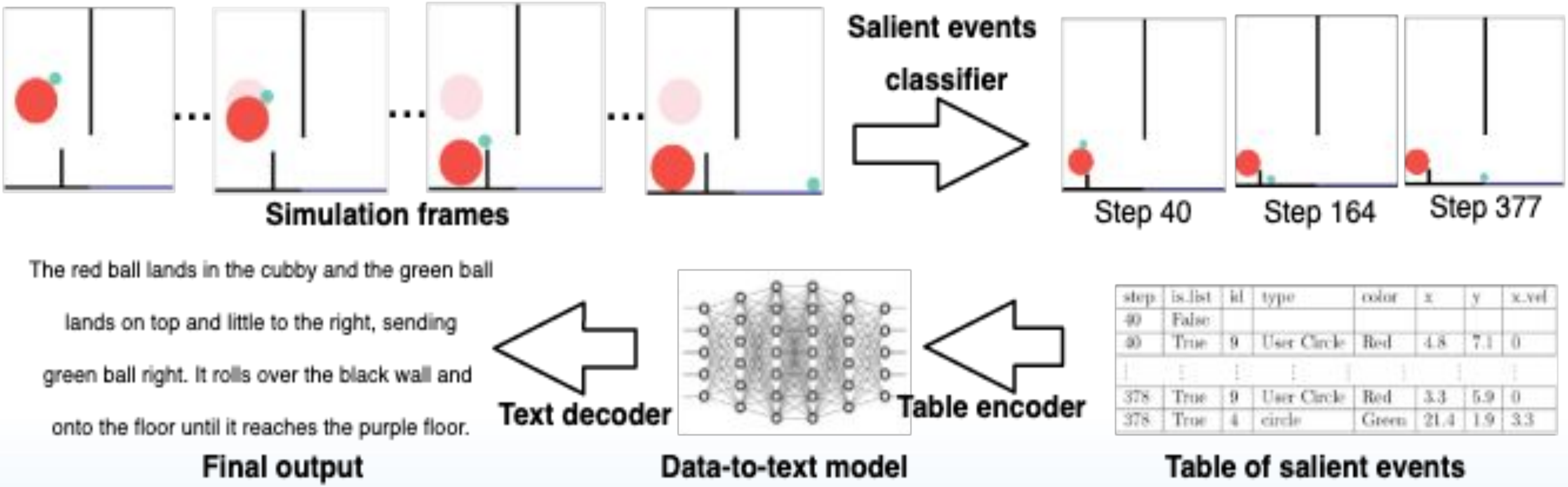


(Puduppully et al., 2019)

Language Model

- GPT (Radford et al., 2018)
- Phase 1: Initial state prompt:
  - "o_1 o_2 o_3 …"
  - Example: "Small dynamic red ball, static grey jar and purple floor"
- Salient collisions prompt:
  - `"<initial state description>. The red ball is placed and"`
  - `Example:` "There is a green vertical bar in grey jar which is placed in the middle of the floor. The floor is purple. The red ball is placed and "
- Not a fair comparison.

Our Framework



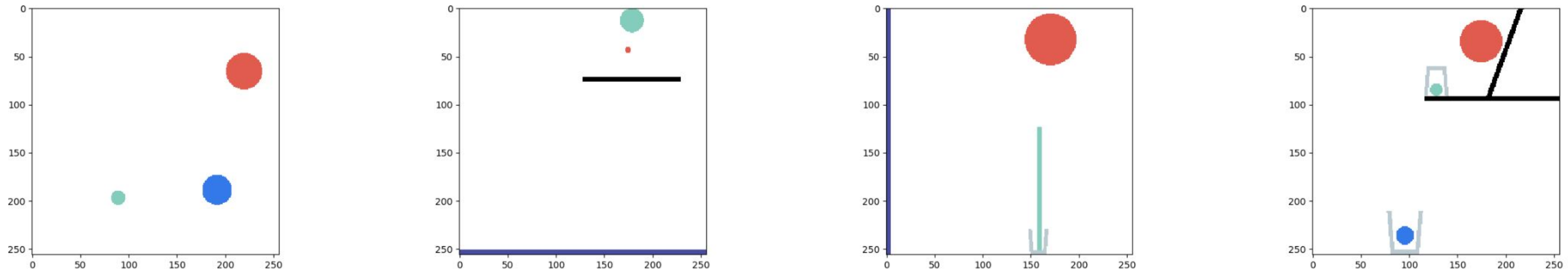
Automatic Evaluations

- Automatic metrics
  - BLEU-1, BLEU-2
  - ROUGE_L
  - METEOR

	Initial State Description				Simulation description			
	BLEU-1	BLEU-2	ROUGE_L	METEOR	BLEU-1	BLEU-2	ROUGE_L	METEOR
GPT (Radford et al., 2018)	18.46	6.35	23.04	9.07	16.06	8.28	24.78	9.91
AVG (Puduppully et al., 2019b)	12.71	8.33	18.20	21.31	19.82	14.44	25.59	25.28
BiLSTM (Puduppully et al., 2019b)	12.62	8.31	17.85	20.93	19.42	13.81	24.72	24.67

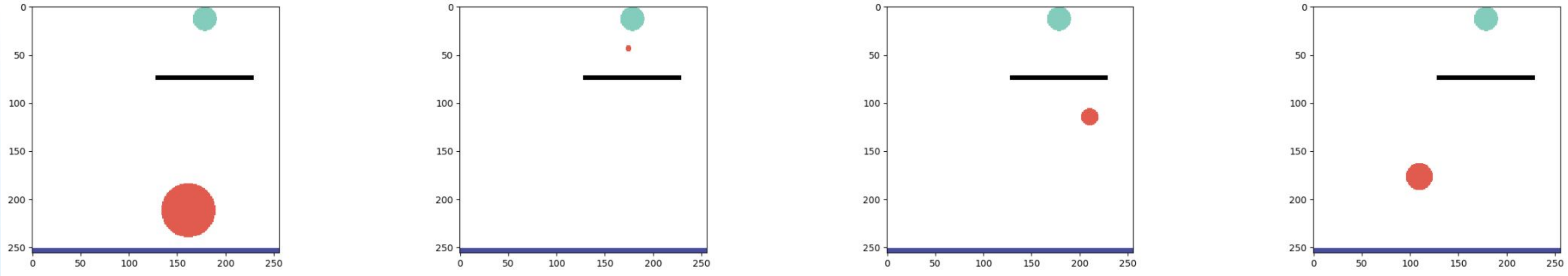
Validity

- **Initial state description**
  - Given: generated description and frame from the simulation along with 3 distractor frames from other simulations



Initial State 1 | Initial State 2 | Initial State 3 | Initial State 4

- **Salient collision description**
  - Given: goal, generated description and initial state frame as well as 3 distractor frames obtained from placement of red ball that leads to no solution



Initial State 1 | Initial State 2 | Initial State 3 | Initial State 4

Coverage

- Collision
  - hit
- Friction
  - Rolling
  - slipping
- Gravity
  - Falling
  - Free fall
- Select which concepts are covered in the description and mention words that imply those concepts

Human Evaluations

	Initial state	Simulation
Random classifier	25.0	25.0
GPT (Radford et al., 2018)	14.8	44.4
AVG (Puduppully et al., 2019b)	85.2	74.1
BiLSTM (Puduppully et al., 2019b)	81.5	51.9
Human Annotation	66.7	63.0

Table 4: Human evaluation for *validity* accuracy of initial state and simulation descriptions on test set.

	Gravity	Friction	Collision
GPT (Radford et al., 2018)	3.9	0.0	6.6
AVG (Puduppully et al., 2019b)	100.0	96.1	86.8
BiLSTM (Puduppully et al., 2019b)	100.0	93.4	84.2
Human Annotation	94.7	57.9	51.3

Table 5: Human evaluation for *coverage* accuracy of physical concepts in simulation descriptions on test set.

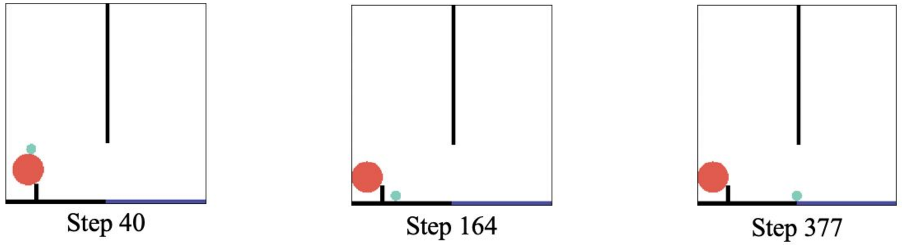
Results

- Top coverage words:
  - Gravity - fall, land, slope, drop
  - Friction - roll, slide, trap, travel, stuck, remain
  - Collision - hit, collide, impact, land, pin, bounce

Table of Simulation Steps

step	is_list	id	type	color	x	y	x_vel	...	is_collision	kind	id_1	type_1	...
40	False							...	True	begin	5	bar	...
40	True	9	User Circle	Red	4.8	7.1	0	...	False				
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
378	True	9	User Circle	Red	3.3	5.9	0	...	False				...
378	True	4	circle	Green	21.4	1.9	3.3	...	False				...

Salient Collision Classification



Data-To-Text Model

Table Encoder → Text Decoder

Final Output

The red ball lands in the cubby and the green ball lands on top and a little to the right, sending the green ball right. It rolls over the short black wall of the cage and onto the floor, where it keeps rolling right towards the purple goal. As a result of impact with the red ball, the green balls moves towards the right, hits the shorter black platform, and continues rolling to the right. It continues rolling until it reaches the purple floor on the right.

Future work

- Train an RL agent to leverage language and perform the task efficiently
  - Reward shaping (Goyal et al., 2019)
  - Generalizing via reading (Zhong et al., ICLR 2020)

ERASER Benchmark for Interpretability in NLP

salesforce

eraserbenchmark.com

Tasks Leaderboard Eraser FAQ Appendix Paper

Eraser

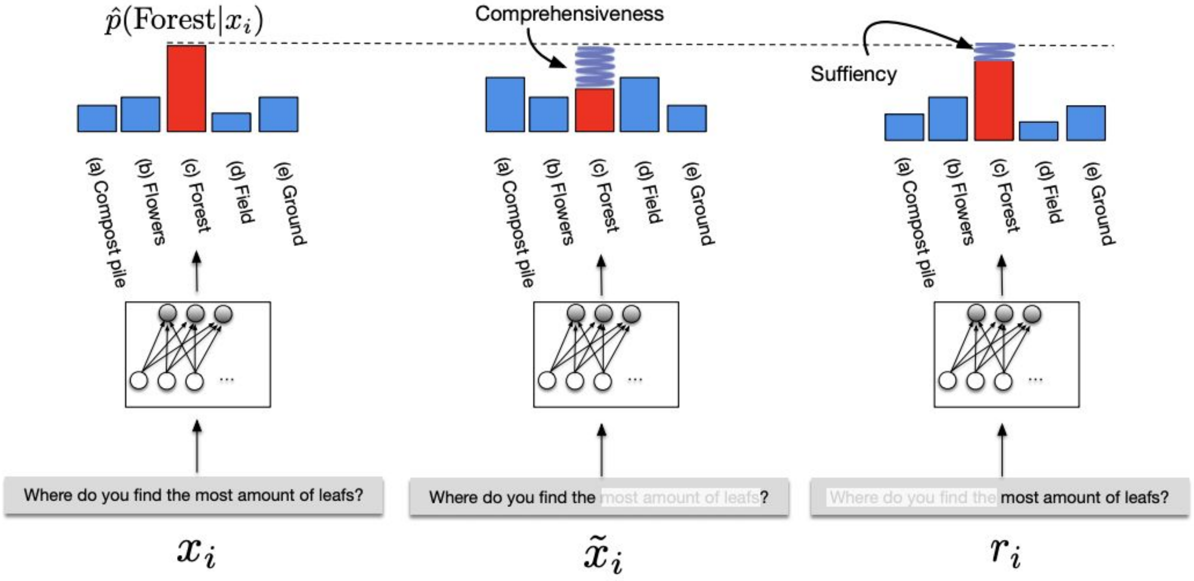
The need for more interpretable models in NLP has become increasingly apparent in recent years. The Evaluating Rationales And Simple English Reasoning (ERASER) benchmark is intended to advance research in this area by providing a diverse set of NLP datasets that contain both document labels and snippets of text marked by annotators as supporting these.

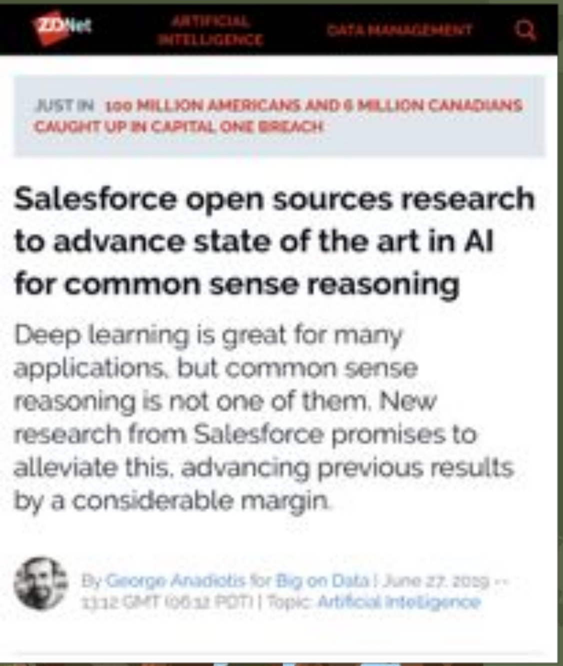
Models that provide rationales supporting predictions can be evaluated using this benchmark using several metrics (*see below*) that aim to quantify different attributes of "interpretability". We do not privilege any one of these, or provide a single number to quantify performance, because we argue that the appropriate metric to gauge the quality of rationales will depend on the task and use-case.

salesforce

Datasets, Models, and Metrics

Dataset	Labels	Instances	Documents	Sentences	Tokens
Evidence Inference	3	9889	2411	156.0	4760.6
BoolQ	2	10671	7030	175.2	3580.1
Movie Reviews	2	2000	1999	36.8	774.1
FEVER	2	110190	4099	12.1	326.5
MultiRC	2	32091	539	14.9	302.5
CoS-E	5	10917	10917	1.0	27.6
e-SNLI	3	568939	944565	1.7	16.0




ZDNet

JUST IN 100 MILLION AMERICANS AND 6 MILLION CANADIANS CAUGHT UP IN CAPITAL ONE BREACH

Salesforce open sources research to advance state of the art in AI for common sense reasoning

Deep learning is great for many applications, but common sense reasoning is not one of them. New research from Salesforce promises to alleviate this, advancing previous results by a considerable margin.

By George Anadiotis for Big on Data | June 27, 2019 -- 13:12 GMT (06:12 PDT) | Topic: Artificial Intelligence

thank you

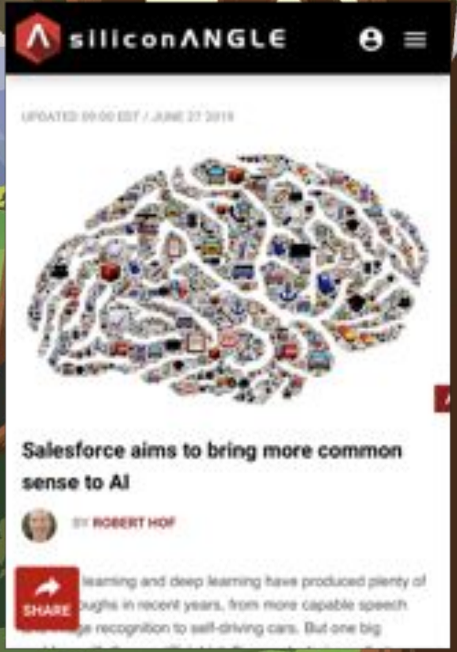
nazneen.rajani@salesforce.com

https://github.com/salesforce/cos-e


datanami

DATA SCIENCE • AI • ADVANCED ANALYTICS

Common Sense Makes Progress with Deep Learning


siliconANGLE

Salesforce aims to bring more common sense to AI

BY ROBERT HOF


VB

Salesforce's AI grasps commonsense reasoning

Kyle Wiggers 1 month ago

Image Credit: Shutterstock

Sophisticated AI models are capable of performing incredible feats, from predicting which patients are likely to develop breast cancer and spotting early signs of glaucoma from eye scans to hallucinating