Commonsense Reasoning in the Wild

Xiang Ren

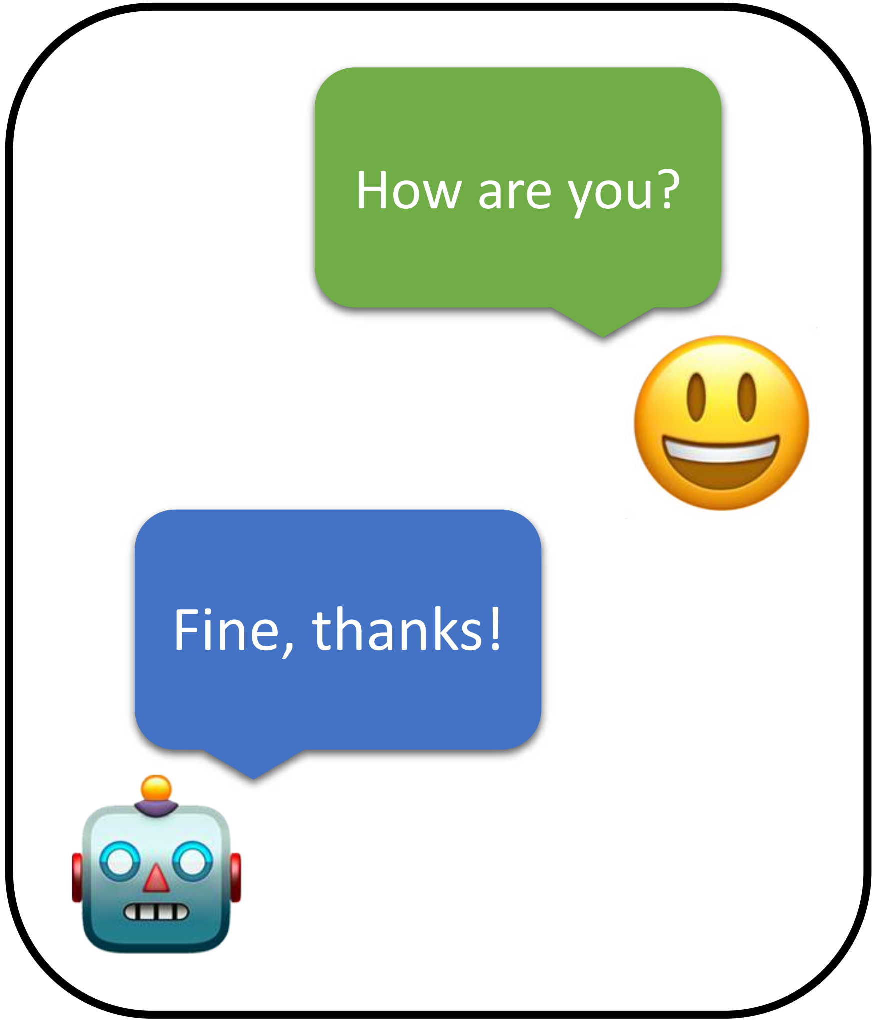
Joint work w/ Bill Yuchen Lin, Pei Zhou, Qinyuan Ye, Jay Pujara,
Yejin Choi, Chandra Bhagavatula, William Cohen

Department of Computer Science & Information Science Institute

University of Southern California

http://inklab.usc.edu

NLP Models on Research Benchmarks

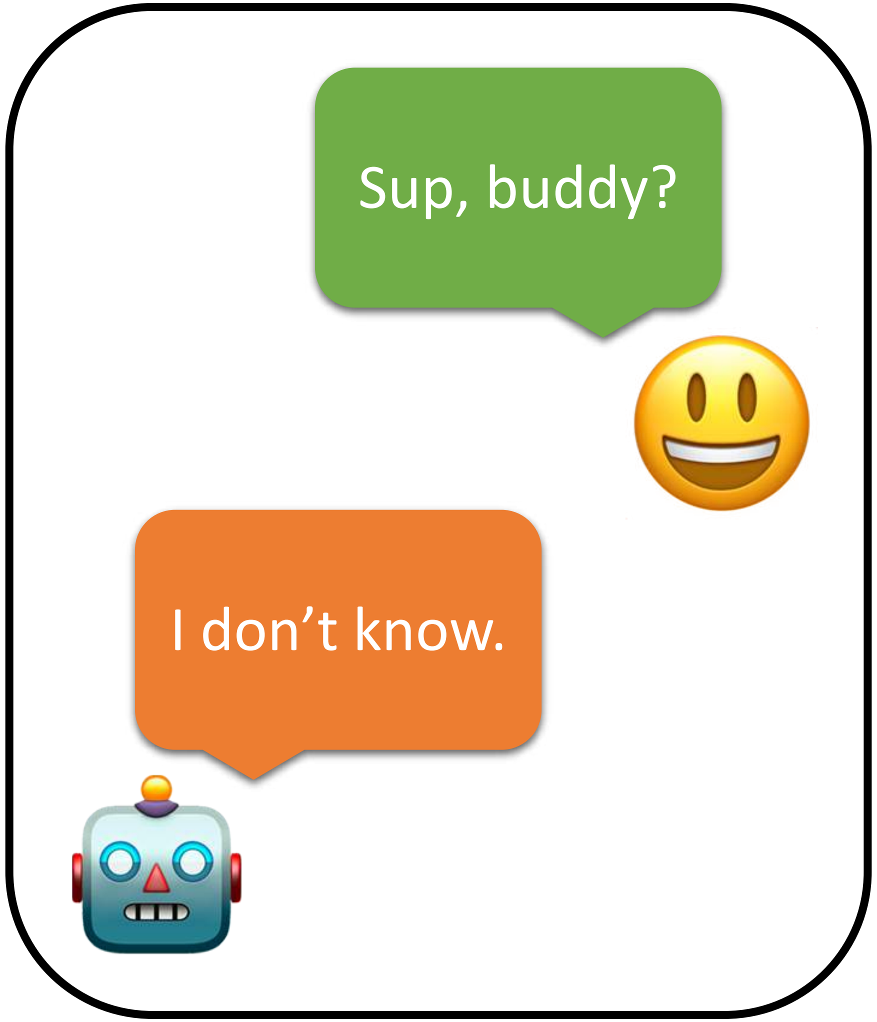


XYZ Leaderboard

Superhuman Performance	90.6	91.0	98.6/99.2	97.4	88.6/63.2	94.7/94.2	92.6	77.4	97.3	68.6	92.7/94.7
	90.4	91.4	95.8/97.6	98.0	88.3/63.0	94.2/93.5	93.0	77.9	96.6	69.1	92.7/91.9
	90.3	90.4	95.7/97.6	98.4	88.2/63.7	94.5/94.1	93.2	77.5	95.9	66.7	93.3/93.8
Human Performance	89.8	89.0	95.8/98.9	100.0	81.8/51.9	91.7/91.3	93.6	80.0	100.0	76.6	99.3/99.7
	89.3	91.2	93.9/96.8	94.8	88.1/63.3	94.1/93.4	92.5	76.9	93.8	65.6	92.7/91.9
	86.7	87.8	94.4/96.0	93.6	84.6/55.1	90.1/89.6	89.1	74.6	93.2	58.0	87.1/74.4
	86.1	88.1	92.4/96.4	91.8	84.6/54.7	89.0/88.3	88.8	74.1	93.2	75.6	98.3/99.2


OUR NLP MODEL ACHIEVED SUPERHUMAN PERFORMANCE ON THE XYZ BENCHMARK
SO, THAT MEANS IT WILL PERFORM WELL ON REAL-WORLD DATA, RIGHT?
imgflip.com
RIGHT?

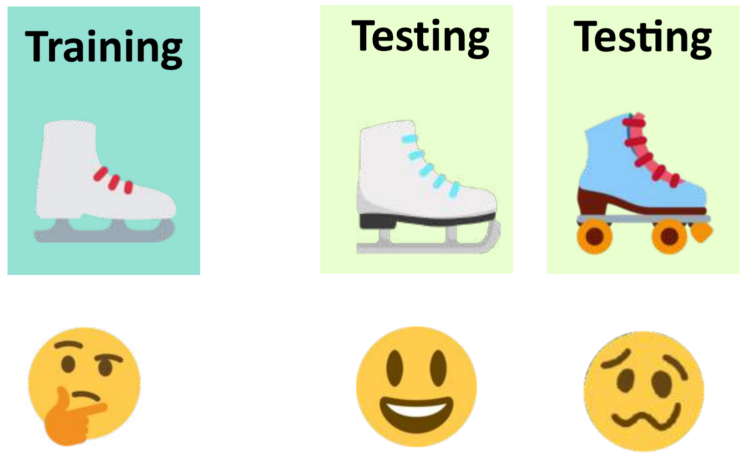
NLP Models in the Wild

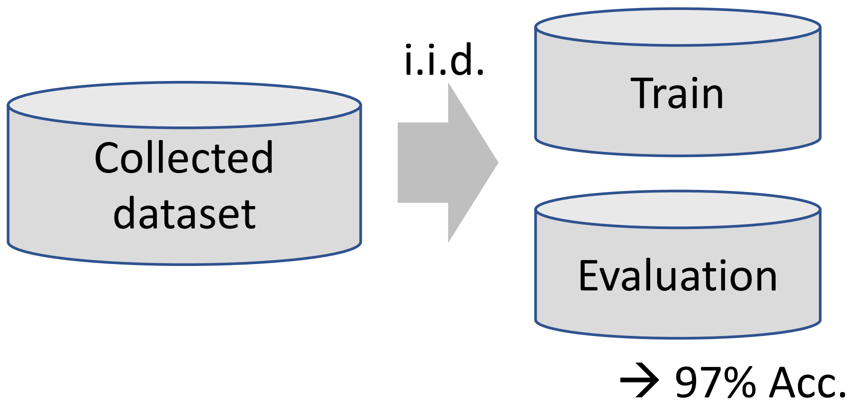


Narrow AI

Performs well **on a specific benchmark**

- Highly customized for narrow tasks
- Hard to deal with unseen situations
- Struggles with under-specified inputs





Narrow AI

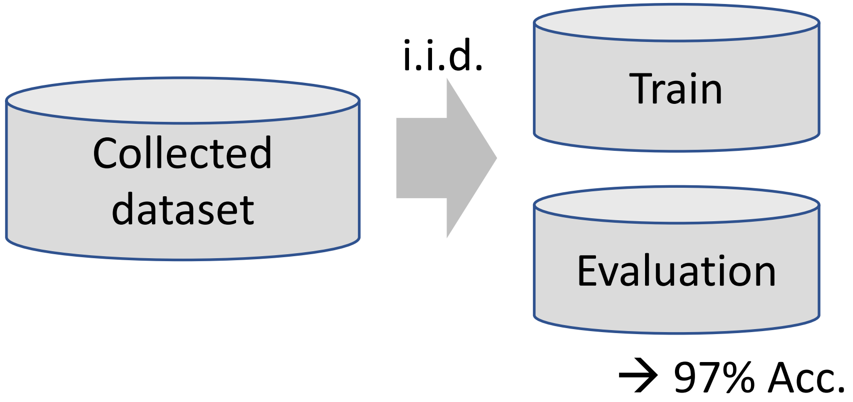
Performs well **on a specific benchmark**

- Highly customized for narrow tasks
- Hard to deal with unseen situations
- Struggles with under-specified inputs

General AI

Performs well **in the real world (in the wild)**

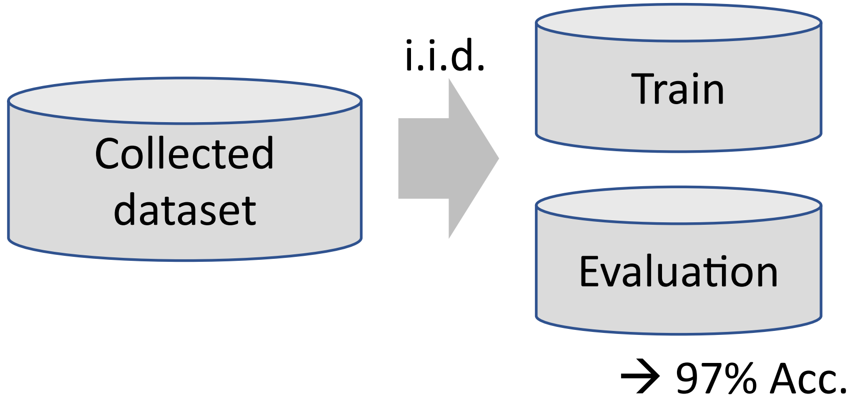
- Applicable to a wide range of tasks
- Generalizes well to novel settings
- Can handle noisy/ambiguous inputs



Narrow AI

Performs well **on a specific benchmark**

- Highly customized for narrow tasks
- Hard to deal with unseen situations
- Struggles with under-specified inputs

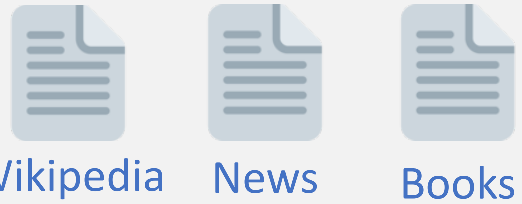


General AI

Performs well **in the real world (in the wild)**

- Applicable to a wide range of tasks
- Generalizes well to novel settings
- Can handle noisy/ambiguous inputs

Generalize to unseen cases



Robust to perturbations

When is the time chage? | Search

Do you mean _when is the time change?_

Training/data efficiency

Trustworthy

(100 years later…)
When was Tokyo 2020 Olympics?

July 2021

What??? Why???

And more…

Narrow AI

Performs well **on a specific benchmark**

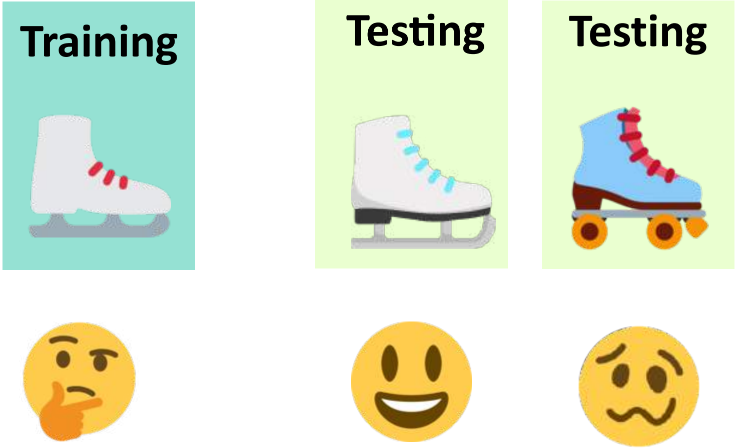
- Highly customized for narrow tasks
- Hard to deal with unseen situations
- Struggles with under-specified inputs

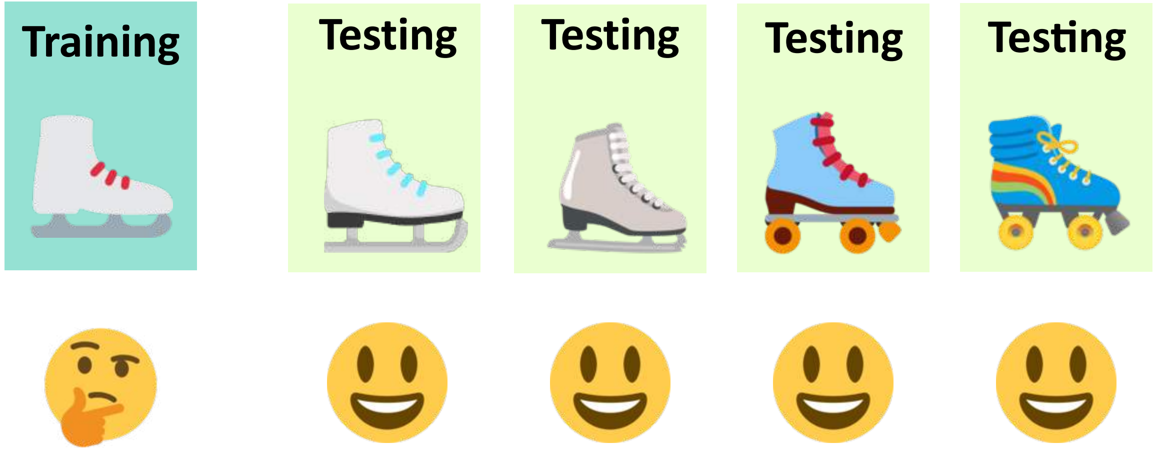
General AI

Performs well **in the real world (in the wild)**

- Applicable to a wide range of tasks
- Generalizes well to novel settings
- Can handle noisy/ambiguous inputs

Commonsense Reasoning!





Why teaching machines common sense?

The human-like ability to understand and generate everyday scenarios (situations, events)

Image source: WikiHow

Why teaching machines common sense?

Common sense -> desirable *inductive bias* for machines to generalize to real-world settings

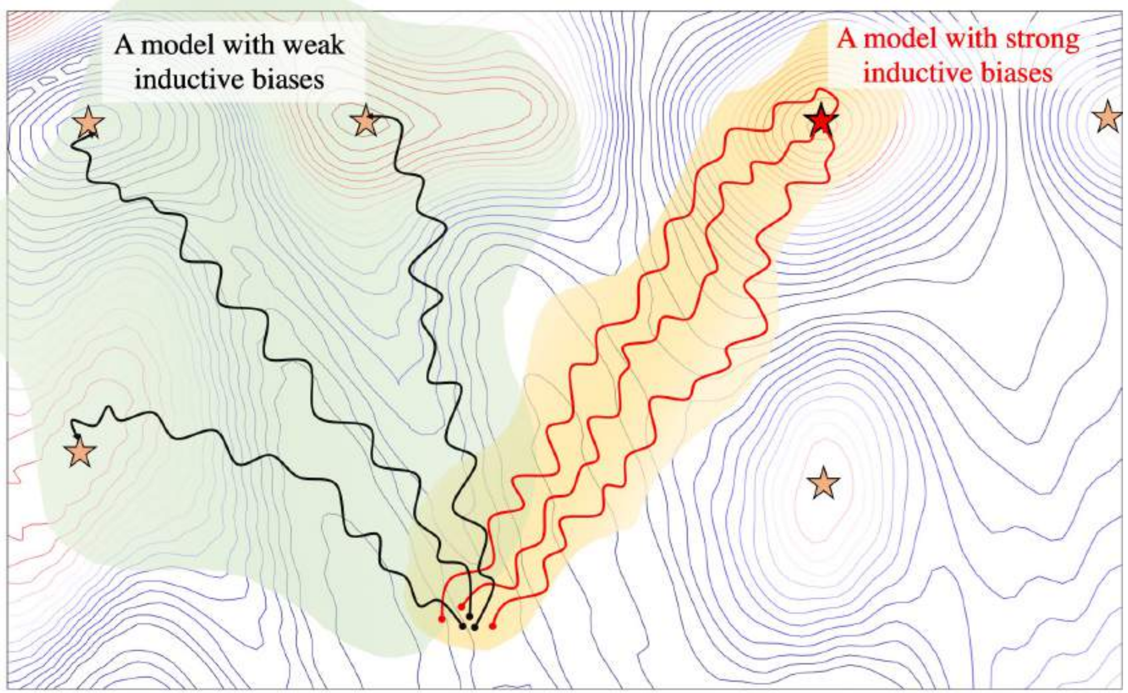
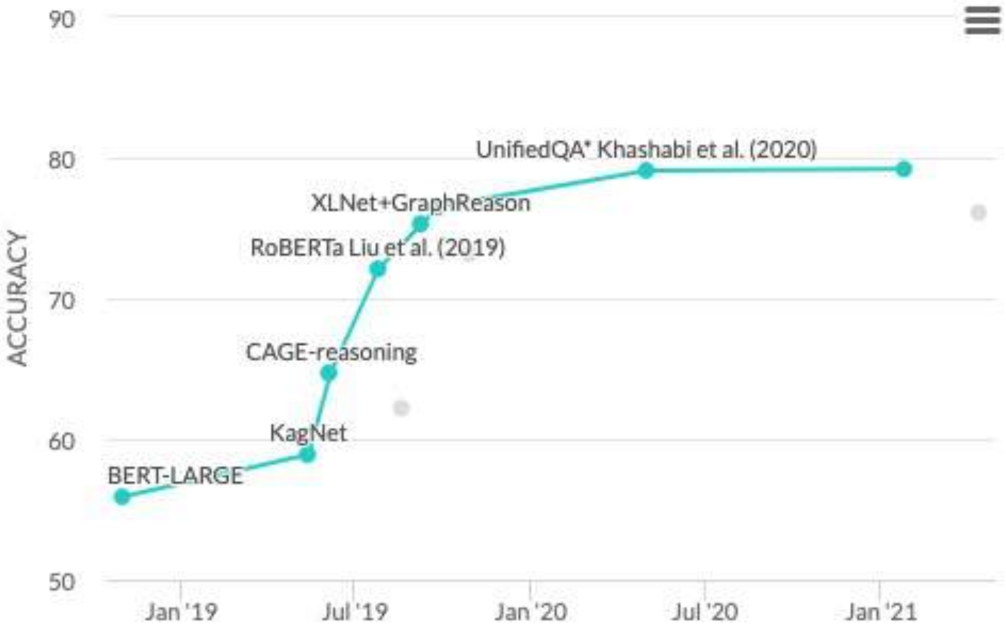


Image source: [WikiHow]

Solving a Commonsense Reasoning Dataset

Goal: Perform well on a test set?

Commonsense Question Answering

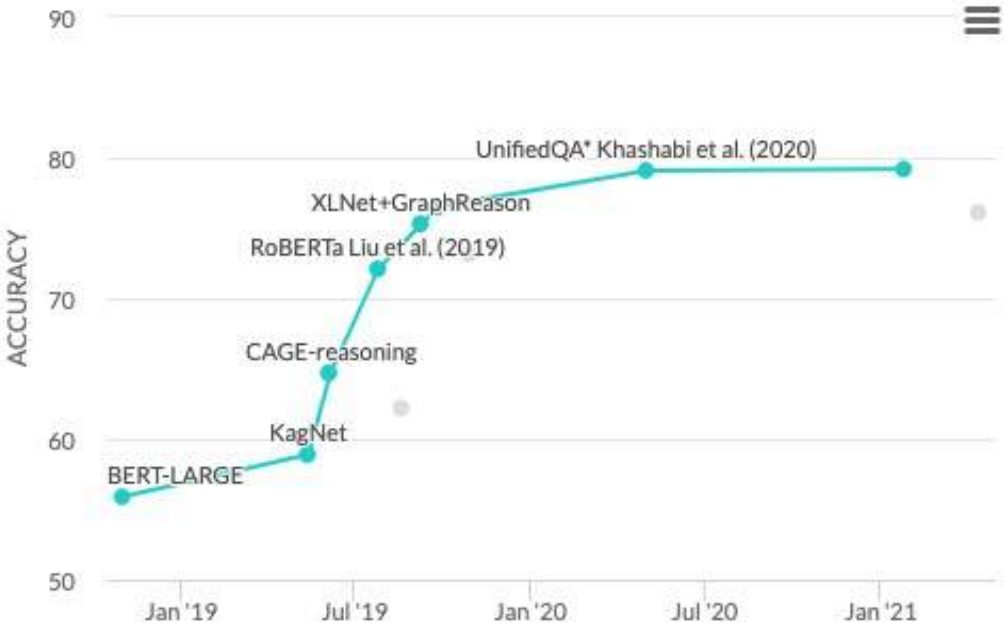


Paper With Code: CommonsenseQA 1.1

Solving a Commonsense Reasoning Dataset

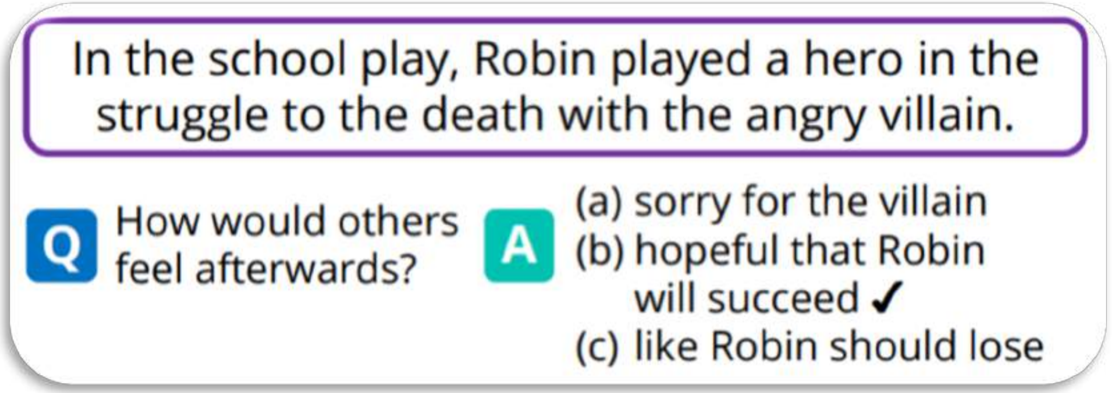
Goal: Perform well on a test set?

Commonsense Question Answering



Paper With Code: CommonsenseQA 1.1

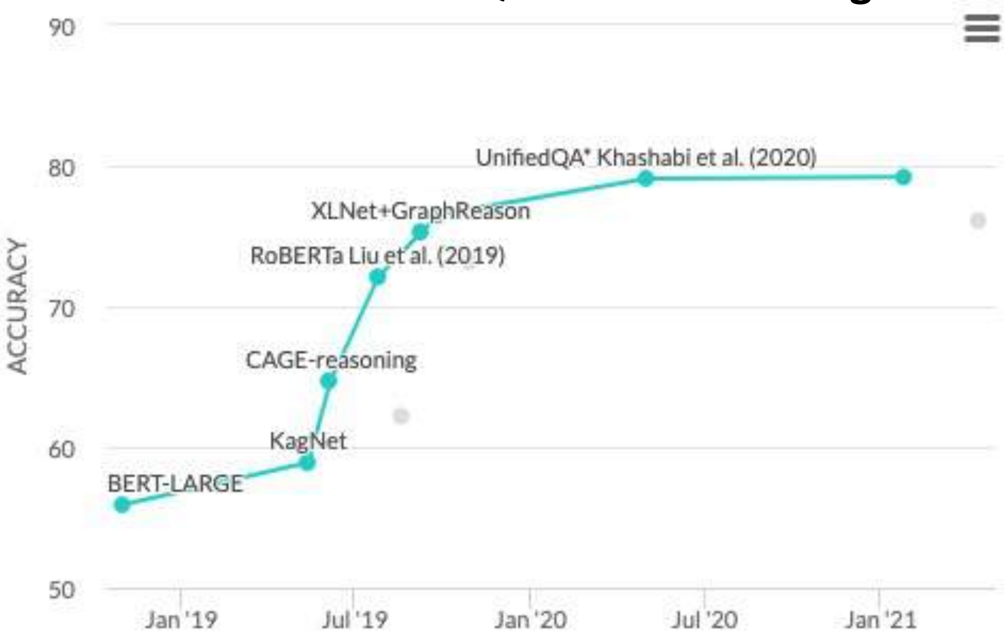
- **discriminative** (closed-ended) reasoning



Solving a Commonsense Reasoning Dataset

Goal: Perform well on a test set?

Commonsense Question Answering



Paper With Code: CommonsenseQA 1.1

- **discriminative** (closed-ended) reasoning
- **logically robust** to linguistic variations

Apples *and* ***oranges*** *grow on* ***trees***

Oranges *and* ***apples*** *grow on* ***trees***

Fruits *grow on* ***trees***

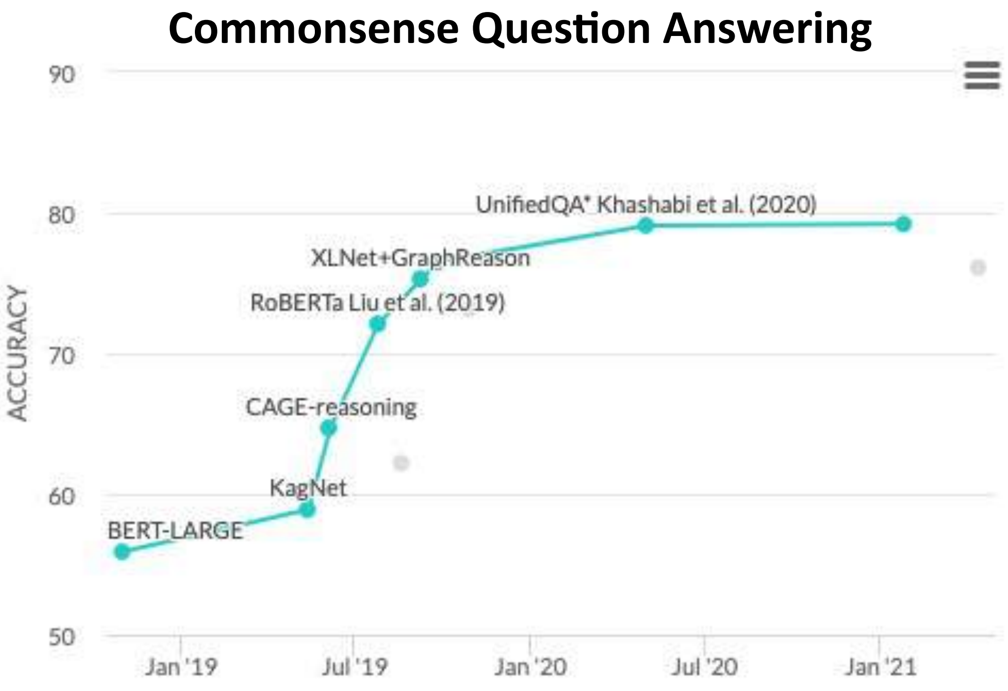
Apples *and* ***oranges*** *grow on* ***plants***

Trees *grow on* ***apples***

Apples *and* ***trees*** *grow on* ***oranges***

Solving a Commonsense Reasoning Dataset

Goal: Perform well on a test set?



Paper With Code: CommonsenseQA 1.1

- **Discriminative** (closed-ended) reasoning
- ***Not* logically robust** to linguistic variations
- ***Not*** quickly **adapt to unseen tasks**

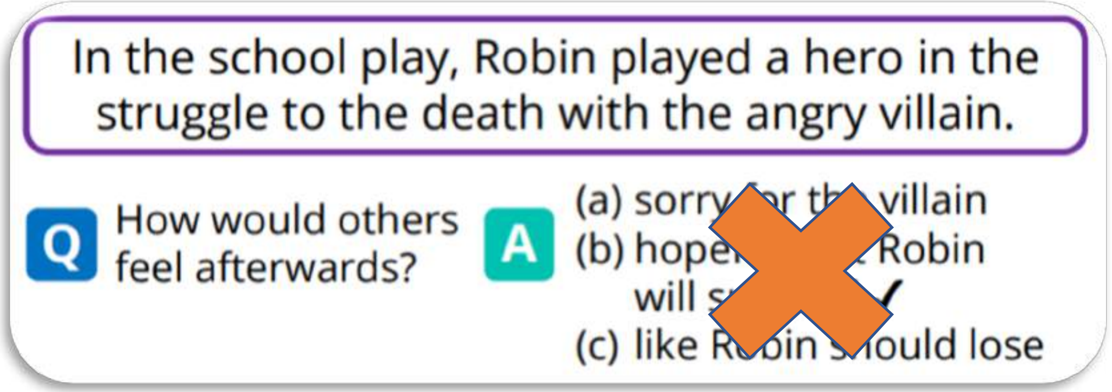
Directly Learning Calculus…

Machine: ? ? ?
Baby: ? ? ?

This talk

This talk

- New ways of formulating CSR challenges:
  - Capable of **open-ended** reasoning



This talk

- New ways of formulating CSR challenges:
  - Capable of **open-ended** reasoning
  - Reason in a **logically consistent** manner

Apples** and **oranges** grow on **trees

Oranges** and **apples** grow on **trees

Fruits** grow on **trees

Apples** and **oranges** grow on **plants

~~***Trees** grow on **apples***~~

~~***Apples** and **trees** grow on **oranges***~~

This talk

- New ways of formulating CSR challenges:
  - Capable of **open-ended** reasoning
  - Reason in a **logically consistent** manner
  - Better at **cross-task generalization**

Directly Learning Calculus...

Machine: ? ? ?
Baby: ? ? ?

Learning Math, History, Geography, Chemistry, ... in high school

Learning Calculus in undergrad
Student: **Yes I can do it!**

CommonGen: A Constrained Text Generation Challenge for Generative Commonsense Reasoning

Bill Yuchen Lin[♥] Wangchunshu Zhou[♥] Ming Shen[♥] Pei Zhou[♥]
Chandra Bhagavatula[♠] Yejin Choi[♠♦] Xiang Ren[♥]
[♥]University of Southern California [♠]Allen Institute for Artificial Intelligence
[♦]Paul G. Allen School of Computer Science & Engineering, University of Washington

(Findings of EMNLP 2020)

What is CommonGen?

- Most current tasks for machine commonsense focus on discriminative reasoning.
  - CommonsenseQA, SWAG.

- Humans not only use **commonsense knowledge** for understanding text, but also for **generating sentences**.

Input:

-A set of common concepts (actions & objects)

Output:

-A sentence that describes an everyday scenario the given concepts.

Why is generative CSR hard?

(1) Relational knowledge are **latent** and **compositional**.

Category	Relations	1-hop	2-hop
Spatial knowledge	AtLocation, LocatedNear	9.40%	39.31%
Object properties	UsedFor,CapableOf,PartOf, ReceivesAction,MadeOf, FormOf, HasProperty,HasA	9.60%	44.04%
Human behaviors	CausesDesire,MotivatedBy, Desires,NotDesires,Manner	4.60%	19.59%
Temporal knowledge	Subevent, Prerequisite, First/Last-Subevent	1.50%	24.03%
General	RelatedTo, Synonym, DistinctFrom, IsA, HasContext,SimilarTo	74.89%	69.65%

Why is generative CSR hard?

(2) Compositional Generalization for unseen concept compounds.

Case Study

Concept-Set: { hand, sink, wash, soap }

[bRNN-CopyNet]: a hand works in the sink .

[MeanPooling-CopyNet]: the hand of a sink being washed up

[ConstLeven]: a hand strikes a sink to wash from his soap.

[GPT-2]: hands washing soap on the sink.

[BERT-Gen]: a woman washes her hands with a sink of soaps.

[UniLM]: hands washing soap in the sink

[BART]: a man is washing his hands in a sink with soap and washing them with hand soap.

[T5]: hand washed with soap in a sink.

1. A girl is washing her hands with soap in the bathroom sink.
2. I will wash each hand thoroughly with soap while at the sink.
3. The child washed his hands in the sink with soap.
4. A woman washes her hands with hand soap in a sink.
5. The girl uses soap to wash her hands at the sink.

Experimental Results

Model \ Metrics	ROUGE-2/L		BLEU-3/4		METEOR	CIDEr	SPICE	Coverage
bRNN-CopyNet (Gu et al., 2016)	7.61	27.79	10.70	5.70	15.80	4.79	15.00	51.15
Trans-CopyNet	8.78	28.08	11.90	7.10	15.50	4.61	14.60	49.06
MeanPooling-CopyNet	9.66	31.14	10.70	6.10	16.40	5.06	17.20	55.70
LevenTrans. (Gu et al., 2019)	10.58	32.23	19.70	11.60	20.10	7.54	19.00	63.81
ConstLeven. (Susanto et al., 2020)	11.82	33.04	18.90	10.10	24.20	10.51	22.20	94.51
GPT-2 (Radford et al., 2019)	17.18	39.28	30.70	21.10	26.20	12.15	25.90	79.09
BERT-Gen (Bao et al., 2020)	18.05	40.49	30.40	21.10	27.30	12.49	27.30	86.06
UniLM (Dong et al., 2019)	21.48	**43.87**	38.30	27.70	29.70	14.85	30.20	89.19
UniLM-v2 (Bao et al., 2020)	18.24	40.62	31.30	22.10	28.10	13.10	28.10	89.13
BART (Lewis et al., 2019)	**22.23**	41.98	36.30	26.30	**30.90**	13.92	30.60	**97.35**
T5-Base (Raffel et al., 2019)	14.57	34.55	26.00	16.40	23.00	9.16	22.00	76.67
T5-Large (Raffel et al., 2019)	22.01	42.97	**39.00**	**28.60**	30.10	**14.96**	**31.60**	95.29
Human Performance	48.88	63.79	48.20	44.90	36.20	43.53	63.50	99.31

(1) Seq2seq models

(2) Fine-tuning pre-trained LMs

Our analysis shows that SPICE has the best correlation with human judgments

CommonGen Leaderboard (V1.1)

Rank	Model	BLEU-4	CIDEr	SPICE
1 Jun 09, 2021	**KFCNet** *MSRA and Microsoft Ads* Email Paper (EMNLP'21)	43.619	18.845	33.911
2 May 18, 2021	**KGR^4** *Anonymous (under review)* Email Document (placeholder)	42.818	18.423	33.564
3 Mar 23, 2021	**KFC (v1)** *MSRA and Microsoft Ads* Email Paper (EMNLP'21)	42.453	18.376	33.277
4 April 25, 2021	**R^3-BART** *Anonymous (under review).* Email Document (placeholder)	41.954	17.706	32.961
5 July 1, 2021	**WittGEN + T5-large** *Anonymous (under review)*	38.233	18.036	31.682

:overage
51.15
49.06
55.70
63.81
94.51
79.09
86.06
89.19
89.13
97.35

T5-Base (Raffel et al., 2019)	14.57	34.55	26.00	16.40	23.00	9.16	22.00	76.67
T5-Large (Raffel et al., 2019)	22.01	42.97	**39.00**	**28.60**	30.10	**14.96**	**31.60**	95.29
Human Performance	48.88	63.79	48.20	44.90	36.20	43.53	63.50	99.31

Open-Ended Commonsense Reasoning

Q: What can help alleviate global warming?

Multiple-Choice/Closed CSR

Input: a question + a few choices

A) air conditioner B) fossil fuel

C) **renewable energy** D) carbon dioxide

Open-Ended CSR

Input: a question only

A large text corpus of commonsense **facts**

Carbon dioxide is the major greenhouse gas contributing to global warming .

Trees remove *carbon dioxide* from the atmosphere through photosynthesis .

renewable energy, ***tree***, solar battery, ...

Output: a ranked list of concepts as answers.

?

*Can machines learn to **reason** without answer candidates?*

(Lin et al., with Google Research, NAACL 2021) https://open-csr.github.io/

Smooth communication requires common sense

Text Message:

"I'm going to perform in front of thousands tomorrow…"

Explicit Knowledge:

Friend is going to perform in front of many people tomorrow

Commonsense Axiom:

Performing in front of people can cause anxiety

Smooth communication requires common sense

Text Message:

"I'm going to perform in front of thousands tomorrow…"

Explicit Knowledge:

Friend is going to perform in front of many people tomorrow

Commonsense Axiom:

Performing in front of people can cause anxiety

Smooth communication requires common sense

Text Message:

"I'm going to perform in front of thousands tomorrow…"

Explicit Knowledge:

Friend is going to perform in front of many people tomorrow

Commonsense Axiom:

Performing in front of people can cause anxiety

Smooth communication requires common sense

Smooth communication requires common sense

Two key challenges

Inference making requires *implicit* commonsense reasoning

Humans fluidly adapt to *diverse* linguistic expressions

RICA: Evaluating Robust Inference Capabilities Based on Commonsense Axioms

Pei Zhou, Rahul Khanna, Seyeon Lee, Bill Yuchen Lin, Daniel Ho,Jay Pujara, Xiang Ren

EMNLP 2021

The RICA Challenge

Define logical primitives

Mine common sense

Represent commonsense in logic

Perturb and convert logic to text

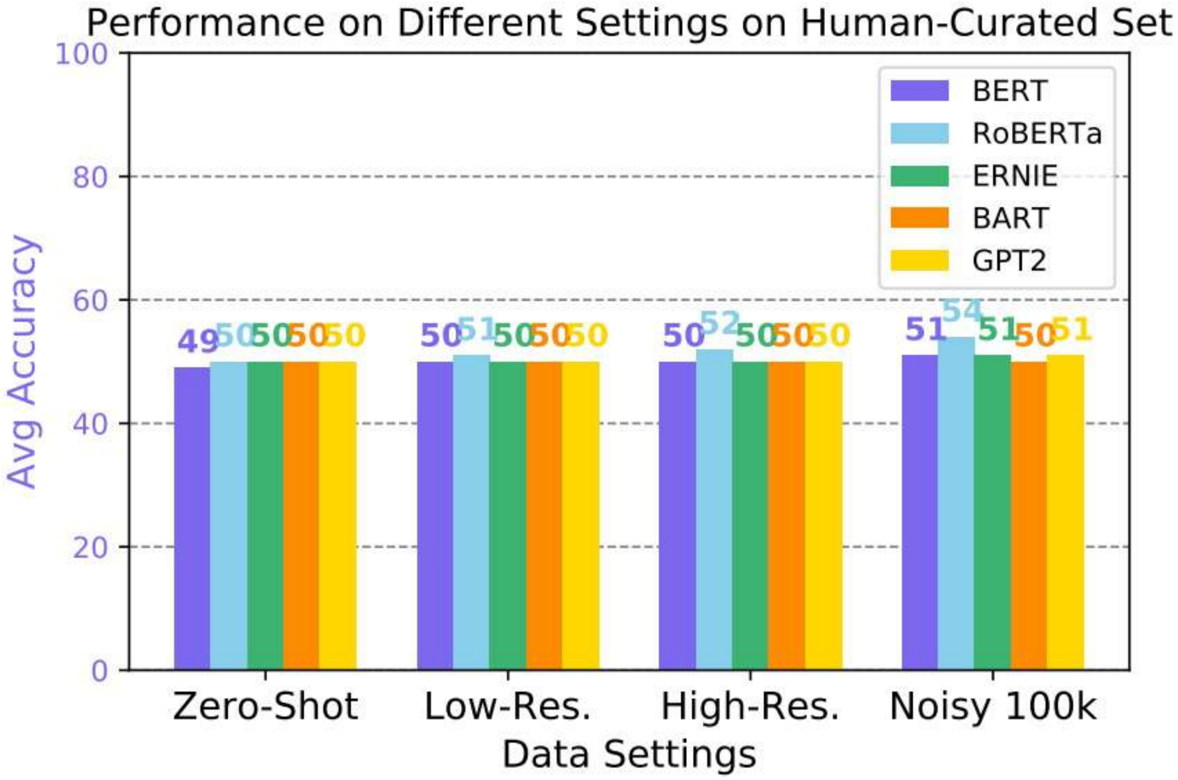
Create commonsense statements that can be used to probe language models

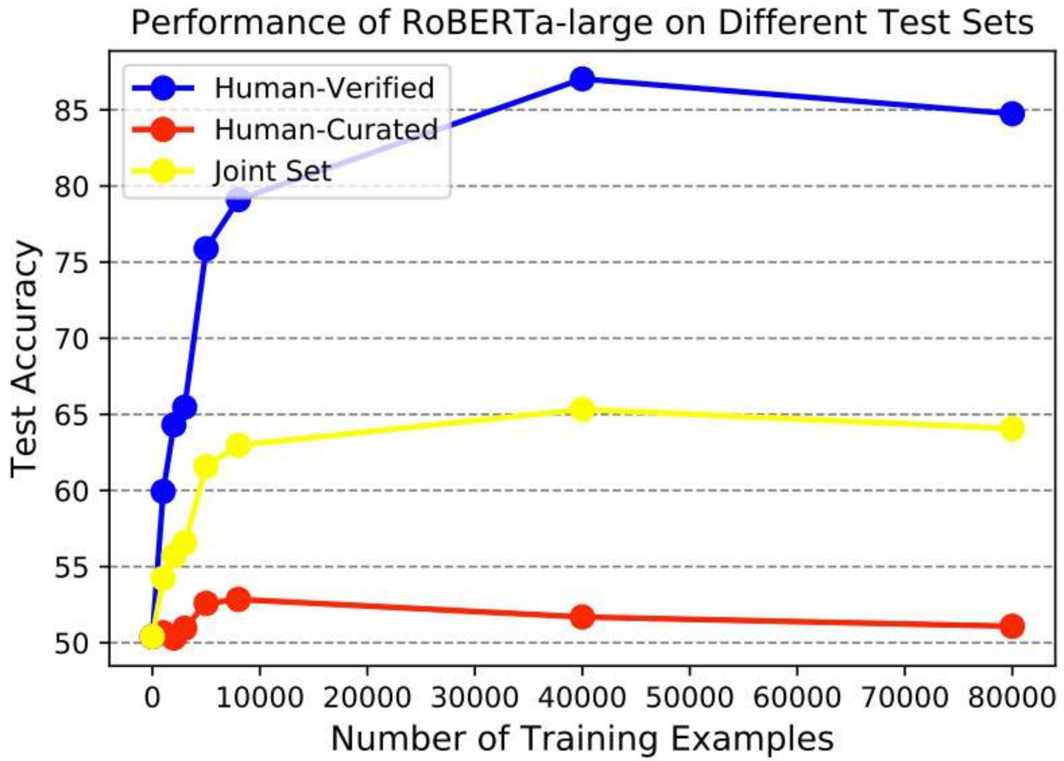
Results: **random guessing, heavy bias, and not robust**

Results: random guessing, heavy bias, and not robust

- Random-guessing like performance for zero-shot and low-resource for all models. Novel entities do not hinder performance.
- More data helps on human-verified set

Human Performance: 91.7%





Cross-task generalization in NLP

- Humans can learn a new task ***efficiently*** with only few examples, by leveraging their knowledge obtained when learning prior tasks.
- We refer to this ability as ***cross-task generalization***.
- How such ability can be ***acquired***, and further ***applied*** to build better few-shot learners across ***diverse NLP tasks***.

Directly Learning Calculus...

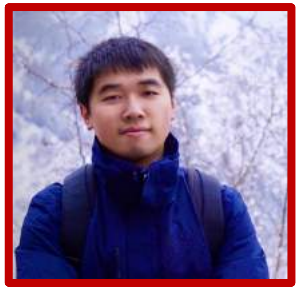
Machine: ? ? ?
Baby: ? ? ?

Learning Math, History, Geography, Chemistry, … in high school

Learning Calculus in undergrad
Student: **Yes I can do it!**

CrossFit 🏋: A Few-shot Learning Challenge for Cross-task Generalization

Qinyuan Ye



Bill Yuchen Lin

Xiang Ren

University of Southern California - Information Sciences Institution

INK Lab @ USC-ISI

inklab.usc.edu

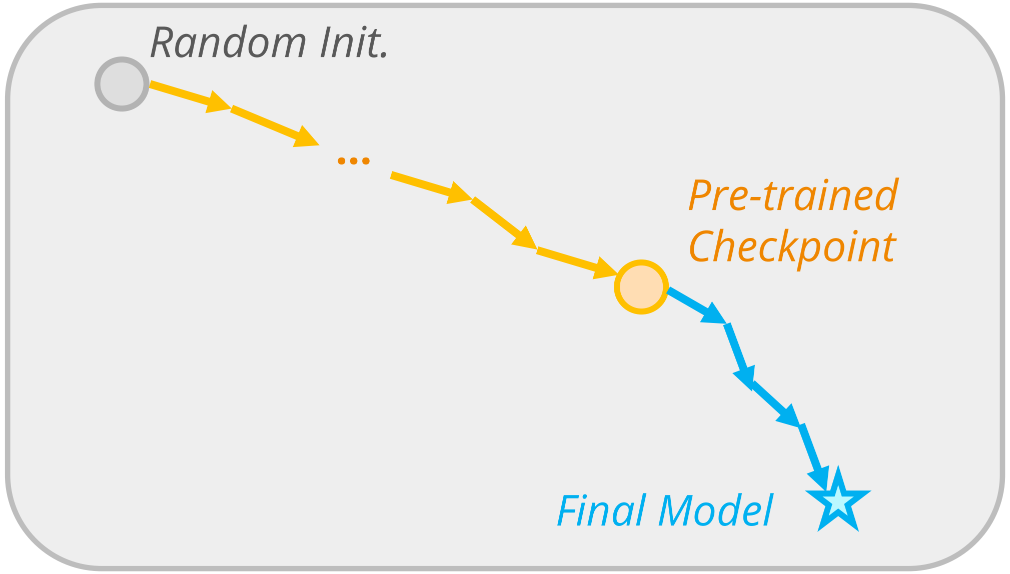
(Ye et al., EMNLP 2021)

Problem Setting

Prevalent Pipeline

Large-scale Pre-training

\+ Downstream Fine-tuning

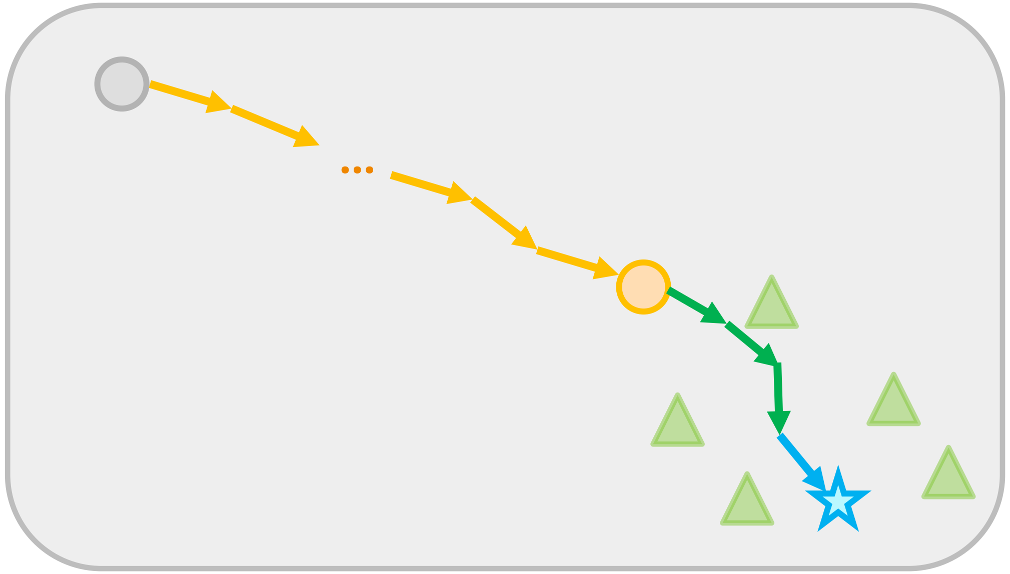


In our CrossFit 🏋️ Setting

Large-scale Pre-training

\+ Upstream Learning on a set of seen tasks

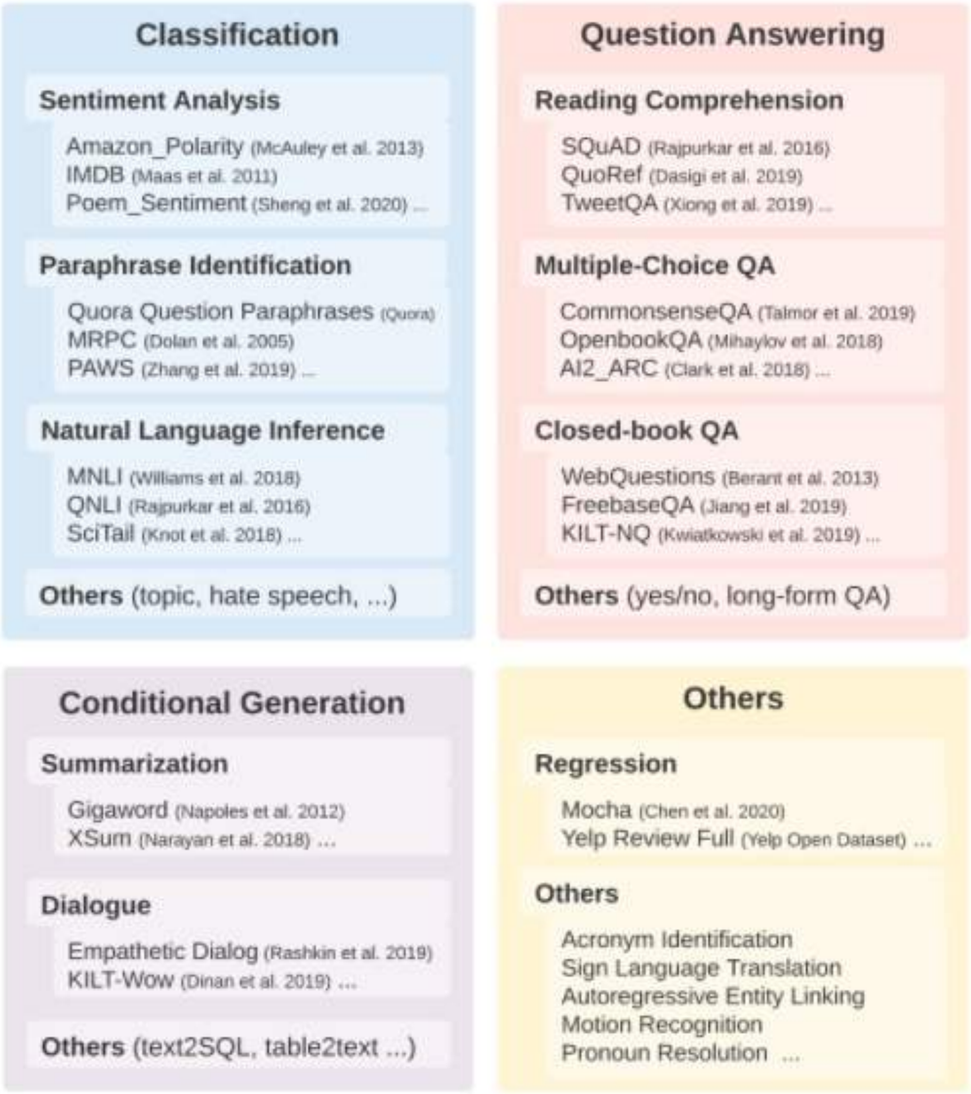
\+ Downstream Fine-tuning on an *unseen* target task



Problem Setting

- To instantiate different settings in **CrossFit** 🏋️ and facilitate in-depth analysis
- We present **NLP Few-shot Gym** 💦, a repository of 160 diverse few-shot NLP tasks.
- We introduce 8 different seen/unseen tasks partitions of these few-shot tasks.

No.	Shorthand	$\mathcal{T}_{train}$	$\mathcal{T}_{dev}$	$\mathcal{T}_{test}$
1	Random	120	20	20
2.1	45cls	45 cls.	10 cls.	10 cls.
2.2	23cls+22non-cls	23 cls. + 22 non-cls.	10 cls.	10 cls.
2.3	45non-cls	45 non-cls.	10 cls.	10 cls.
3.1	Held-out-NLI	57 non-NLI cls.	/	8 NLI
3.2	Held-out-Para	61 non-Paraphrase cls.	/	4 Para. Iden.
4.1	Held-out-MRC	42 non-MRC QA	/	9 MRC
4.2	Held-out-MCQA	29 non-MC QA	/	22 MC QA



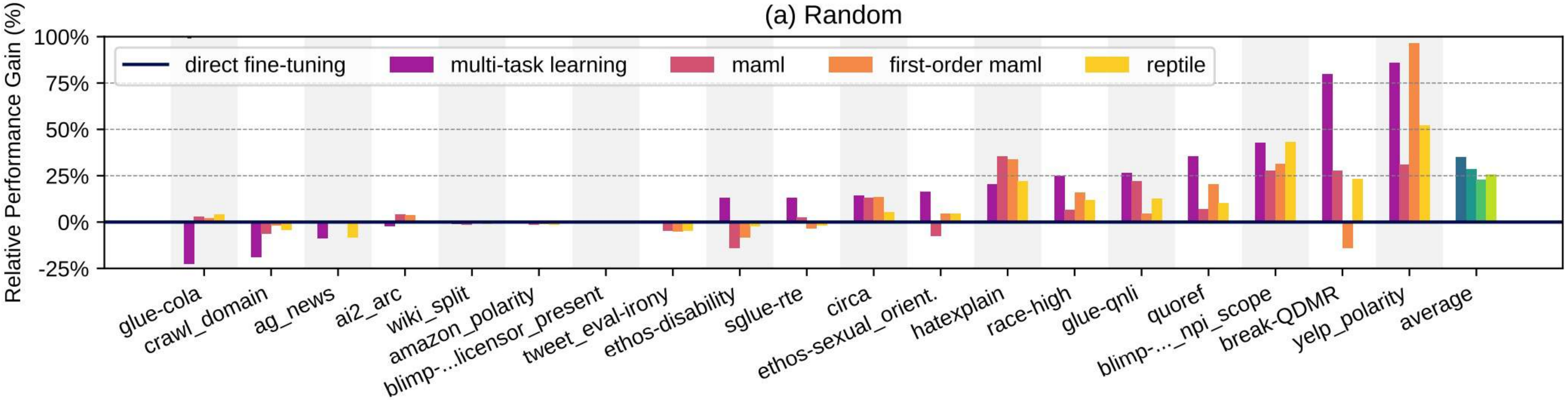
Key Findings

- **Q1. Does upstream learning help cross-task generalization?**

Key Findings

- **Q1. Does upstream learning help cross-task generalization?**
  - We tried applying **multi-task learning** and **meta-learning** methods during the upstream learning stage.

height = relative performance gain after upstream learning



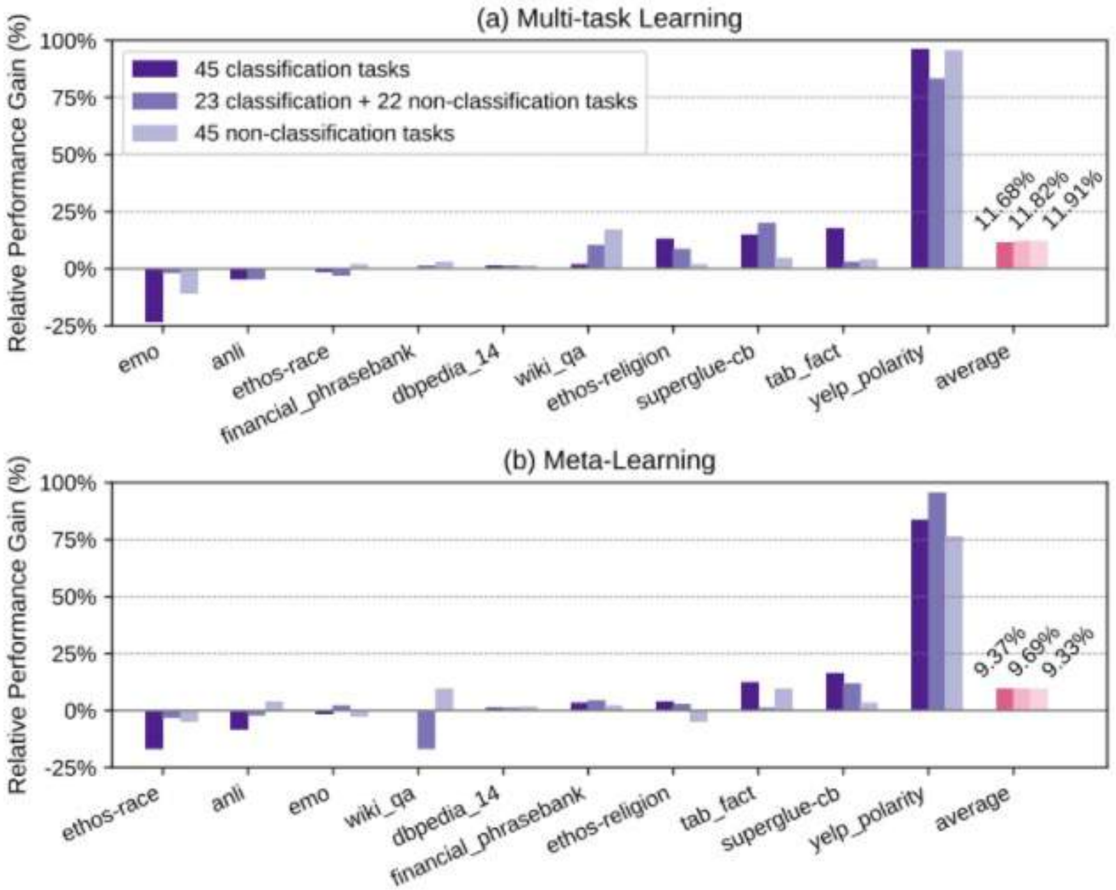
- **Yes!** **These methods do help pre-trained LMs to acquired cross-task generalization.**

Key Findings

- **Q2. “Well-rounded” or “specialized”? How to select tasks during upstream learning?**

- Controlled experiments by fixing the downstream tasks to be 10 classification tasks.
- The upstream tasks are
  - **100% classification tasks**
  - **50% classification + 50% non-classification tasks**
  - **100% non-classification tasks**

- Classification tasks and non-classification tasks seem to be equivalently helpful.
- **Our understanding of tasks may not align with how models learn transferable skills**.

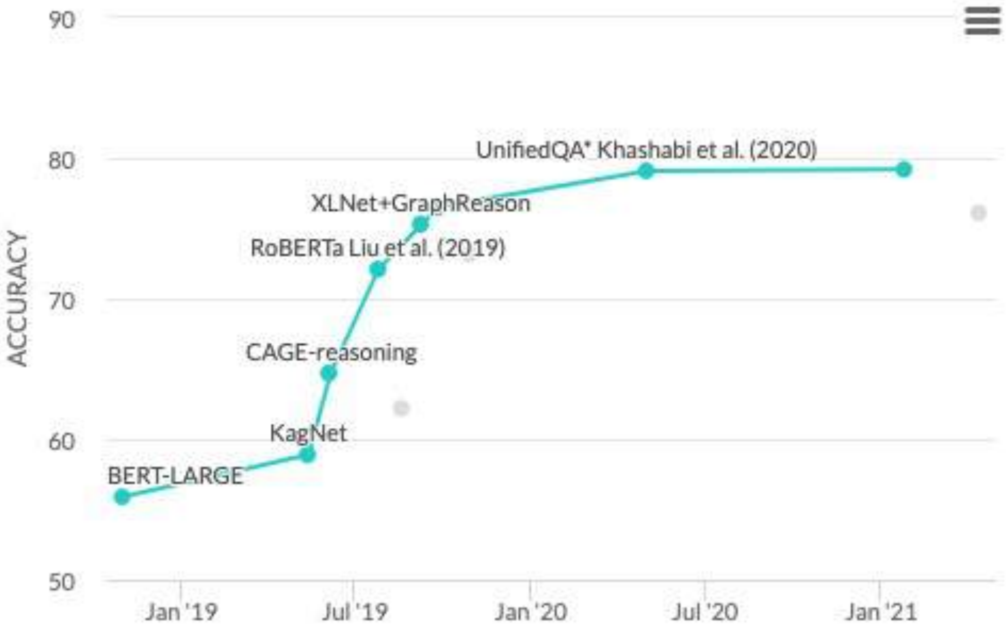


Takeaways

Solving a Commonsense Reasoning Dataset

Goal: Perform well on a test set

Commonsense Question Answering



Paper With Code: CommonsenseQA 1.1

Solving Commonsense Reasoning

Goal: Satisfy the real-world needs

Generalize to unseen cases

Wikipedia News Books

Robust to perturbations

When is the time chage? [Search]

Do you mean *when is the time change?*

Training/inference efficiency

Trustworthy

(100 years later...)
When was Tokyo 2020 Olympics?

July 2021

What??? Why???

And more...