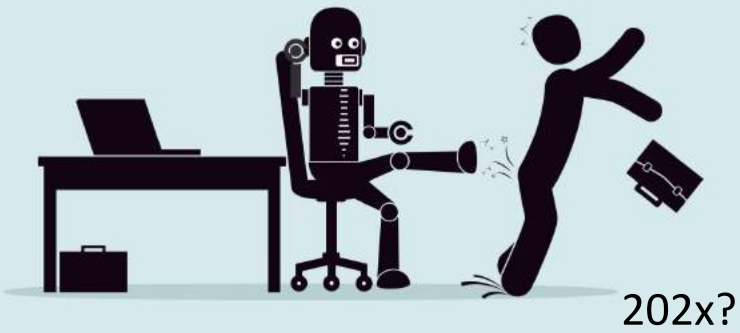
Commonsense Reasoning in the Wild

Xiang Ren

Department of Computer Science
& Information Science Institute
University of Southern California
http://inklab.usc.edu

Super-Human Performance in AI?



Alibaba and Microsoft AI beat human scores on Stanford reading test

Neural networks edged past human scores on the measure of machine reading.

Rob LeFebvre, @roblef
01.15.18 in Personal Computing

10 Comments

937 Shares

2018 — human-level performance on reading comprehension on SQuAD (Stanford QA dataset)

2017 — super-human performance on speech recognition

2016 — Google neural machine translation

2015 — super-human performance on image captioning

2015 — super-human performance on object recognition

Timeline credits: ACL2020 Tutorial on Commonsense

Solving a "dataset" vs the underlaying "task"

Image credit: Yejin Choi

Narrow AI

Highly customized for narrow tasks

Hard to deal with to unseen situations

Struggles with under-specified inputs

"Who was the 16th president of the USA?"

🤗

Image Sources: [We Are Here Japan](), [The Verge]()

Narrow AI

Highly customized for narrow tasks

Hard to deal with to unseen situations

Struggles with under-specified inputs

“what should I dress tonight?”

🤯

Image Sources: [We Are Here Japan](), [The Verge]()

"what should I dress tonight?"

🤗

General AI

Applicable to a wide range of tasks

Generalizes well to novel settings

Can handle noisy/ambiguous inputs

Performs well on a **specific benchmark**

Narrow AI

Highly customized for narrow tasks

Hard to deal with to unseen situations

Struggles with under-specified inputs

→

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

General AI

Applicable to a wide range of tasks

Generalizes well to novel settings

Can handle noisy/ambiguous inputs

Performs well on a **specific benchmark**

Narrow AI

Highly customized for narrow tasks

Hard to deal with to unseen situations

Struggles with under-specified inputs

Commonsense Reasoning!

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

General AI

Applicable to a wide range of tasks

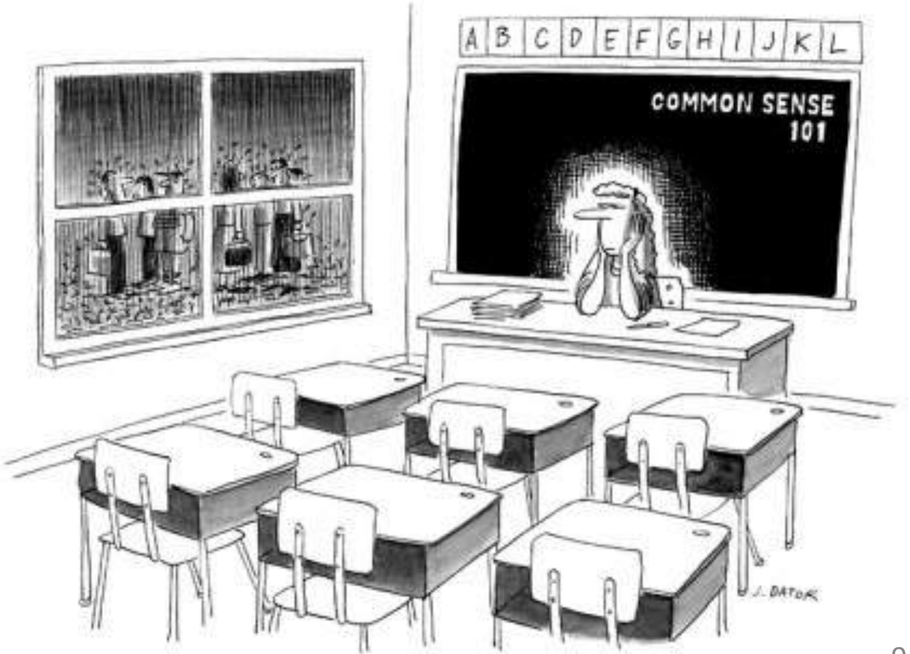
Generalizes well to novel settings

Can handle noisy/ambiguous inputs

Reference: Gunning, 2018

What is common sense?

- Definition of Common Sense: the basic level of practical knowledge and reasoning
  - Physical objects, properties, affordance / temporal, numerical / human behaviors, social norms / commonsense
  - The computation process of manipulating commonsense knowledge to make compositional logical inference



- crucial to functioning well in the real world
- rarely taught explicitly
- yet shared by almost everyone

References: Sap et al., 2020, Shwartz, 2021   Image Source: MIT

Why teaching machines common sense?

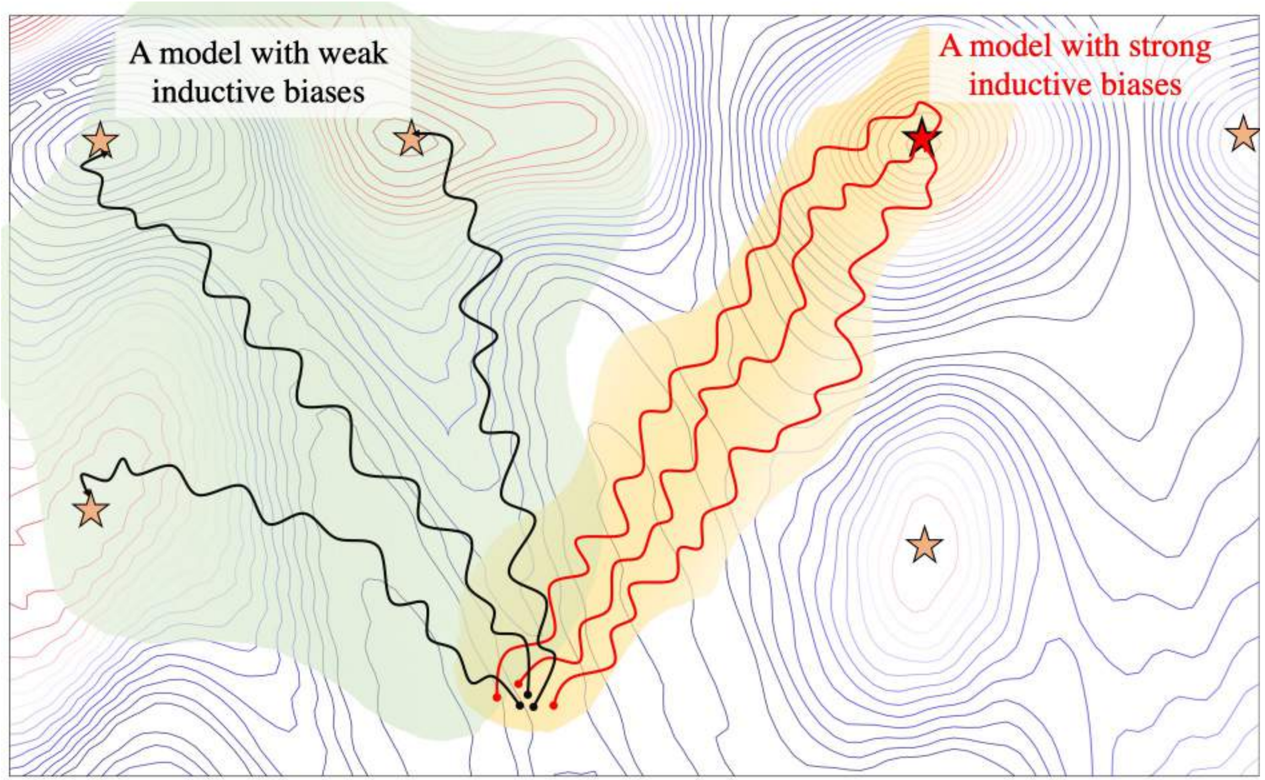
- The human-like ability to understand and generate everyday scenarios (situations, events)

References: Sap et al., 2020, Shwartz, 2021

Image Source: WikiHow

Why teaching machines common sense?

- Common sense = desirable ***inductive bias*** for machines to generalize to real-world settings
  - Avoid learning *spurious patterns* from training data



Reference: [Gunning, 2018]

Image Source: [Samira Abnar]

Common sense is hard for machines to learn!

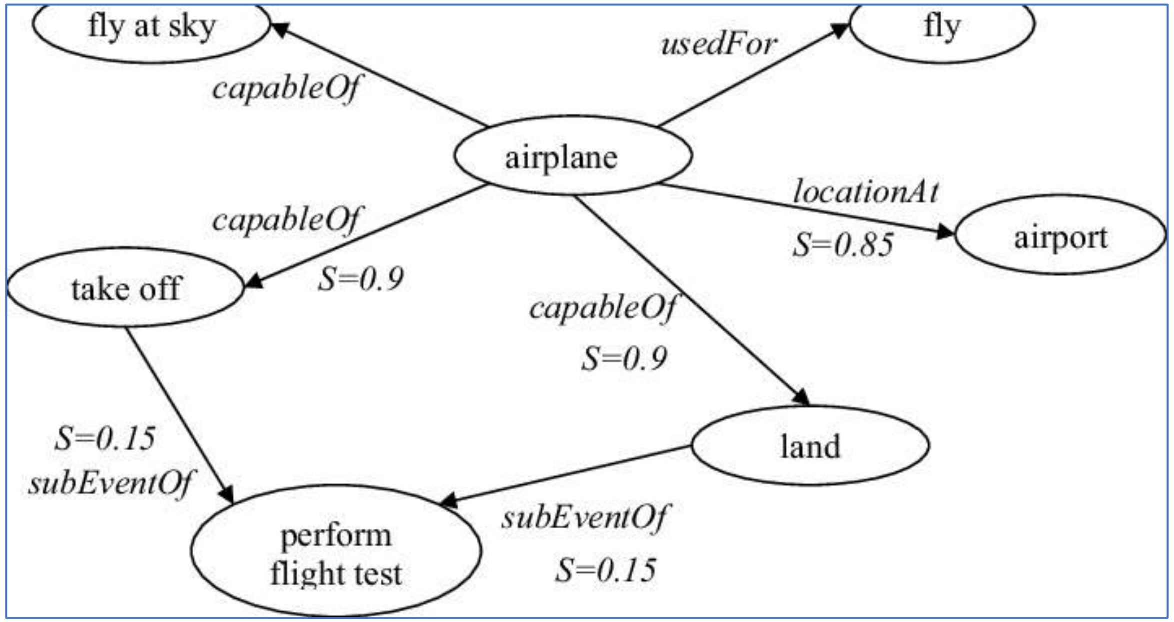
- Humans seldom express commonsense knowledge in natural language
  - Too obvious to even say!
  - e.g., “You might bake a cake because you want people to eat the cake.”

References: [Sap et al., 2020](), [Shwartz, 2021]()

Image Source: [Gemma’s Bigger Bolder Baking]()

Recent Attempts: Neural-symbolic CSR

- Focus on using **knowledge graphs (KGs)** as external information source for CSR tasks
  - KGs provide **abundant commonsense knowledge** beyond text corpora and task inputs
  - Improve generalization by training model to **reason over KGs' symbolic structure**



ConceptNet
An open, multilingual knowledge graph

ATOMIC
An Atlas of Machine Commonsense for If-Then Reasoning

References: Lin et al., 2019, Feng et al., 2020

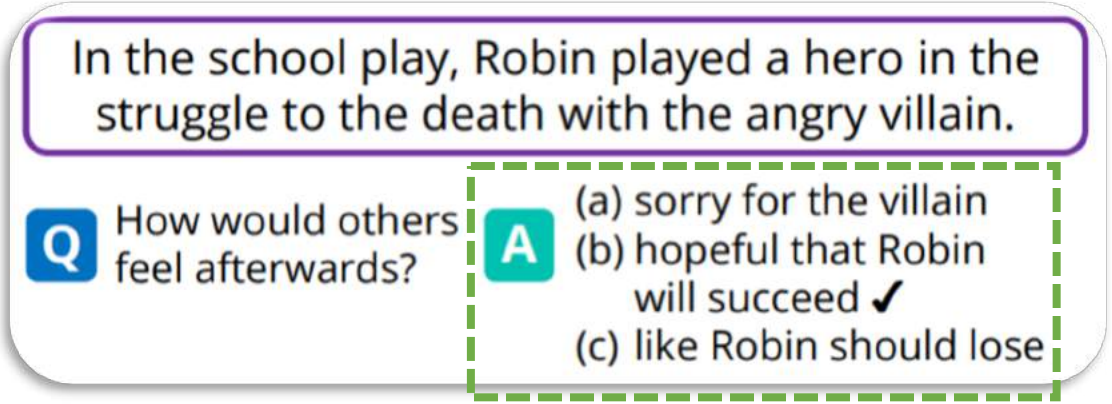
Image Source: Let the Machines Learn

Existing CSR Systems

- Limitations

Existing CSR Systems

- Limitations
  - Designed for **discriminative** (closed-ended) reasoning



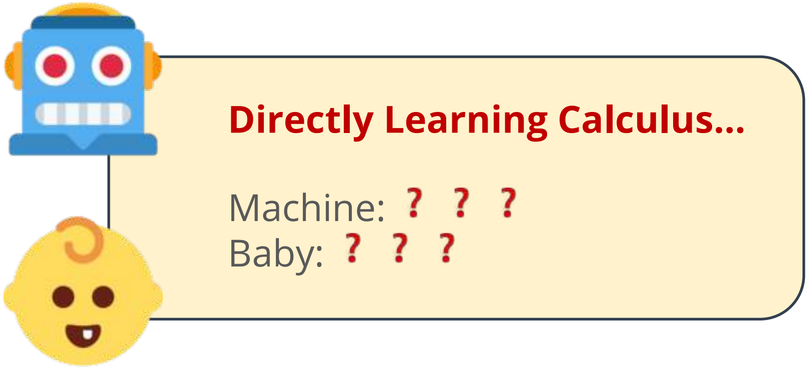
Existing CSR Systems

- Limitations
  - Designed for **discriminative** (closed-ended) reasoning
  - Not **logically robust** to linguistic variation/perturbation

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***

Existing CSR Systems

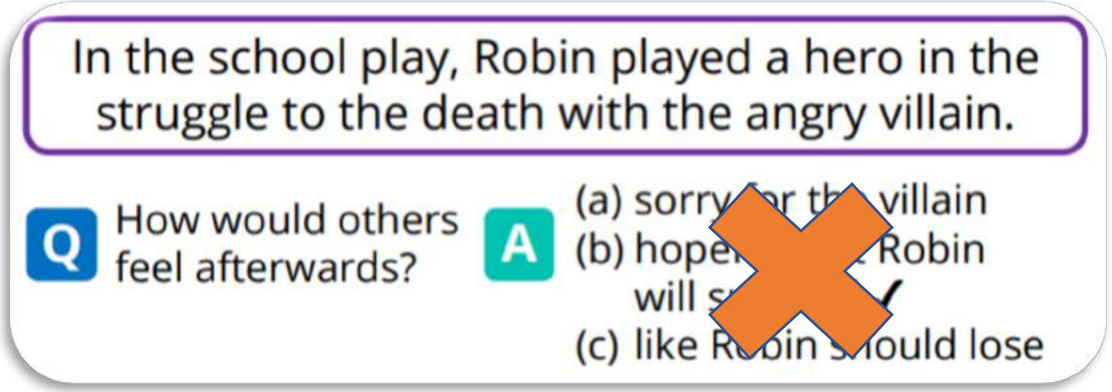
- Limitations
  - Designed for **discriminative** (closed-ended) reasoning
  - Not **logically robust** to linguistic variation/perturbation
  - Don’t easily **adapt to unseen tasks**



Our Contributions

Our Contributions

- Making CSR systems:
  - Capable of **open-ended** reasoning



Our Contributions

- Making CSR systems:
  - 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***~~

Our Contributions

- Making CSR systems:
  - 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!**

[EMNLP 2020] 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

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 CommonGen hard: Two key Challenges

(2) Compositional Generalization for unseen concept compounds.

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

(3) Agreement

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.

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/

Why is OpenCSR challenging?

1) Latent Multi-Hop Structures (vs. factoid questions).

Who voices the *dog* in the TV show *Family Guy* ?

A multi-hop, factoid question from **HotpotQA**.

q_1= the dog in the TV show Family Guy

q_2= who voices [**q_1. answer**]

Clear, **explicit** hints for querying **evident relations** between **named entities**.

What can help alleviate global warming?

q_1= what *contributes to* global warming

q_2= what *removes* [**q_1. answer**]

Latent, **implicit** hints for querying **complex relations** between **concepts**.

2) Very Large Search Space (vs. multiple-choice setting).

3) Much Denser Entity Links (vs. named entities).

DrFact: multi-hop reasoning over fact corpus

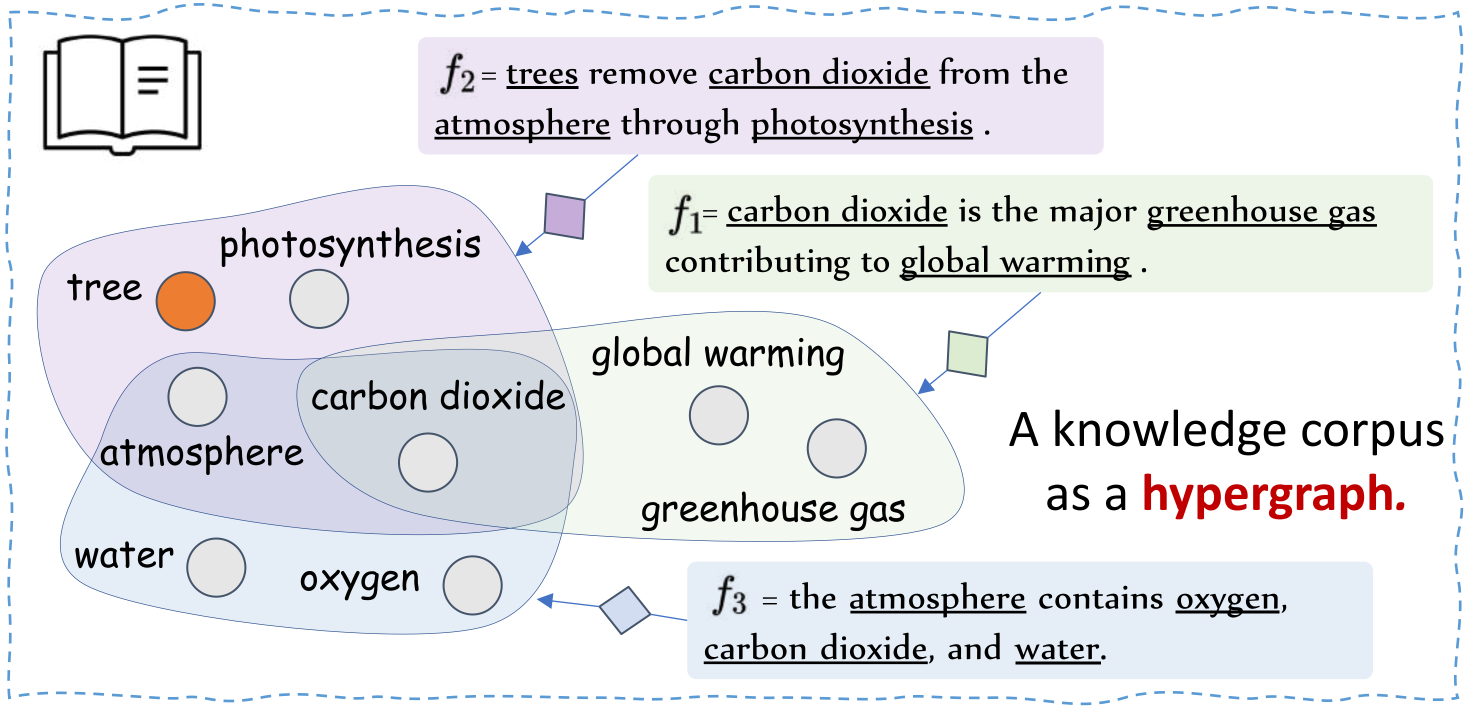
a **corpus** of common-sense facts, e.g., **GenericsKB**. $\mathcal{F}$

$f_i \in \mathcal{F}$

A **fact** is a sentence of generic commonsense knowledge

$c_j \in \mathcal{V}$

A **concept** is a noun or noun-chunk that are mentioned in $\mathcal{F}$



Current Pre-Trained Language Models (PTLMs)

PTLMs

Corpus

... **Copy paper** is thinner than **printer** paper, which doesn't make a huge difference when you're printing text, but it does when you're printing large images. Images require a lot of **ink** and because **copy paper** has a thinner structure, the **ink** will need to spread out more for the **paper** to absorb it all. ...

Pre-train

Text Infilling / MLM

AI2 Allen Institute for AI

UNIFIED-QA

What do you fill with ink to write notes on a piece of copy paper ?

1. Fountain pen

2. Pencil case

3. Printer

4. Notepad

Output:

Prediction [small, 60 million parameters]: pencil case

Prediction [large, 770 million parameters]: printer

Base : pencil case

Large : printer

The model may be **sensitive** to the **co-occurence (ink, copy, paper)**

Pre-training Text-to-Text Transformers for Concept-centric Common Sense (ICLR 2021)

Wangchunshu Zhou*, **Dong-Ho Lee***, Ravi Kiran Selvam,
Seyeon Lee, Bill Yuchen Lin, Xiang Ren

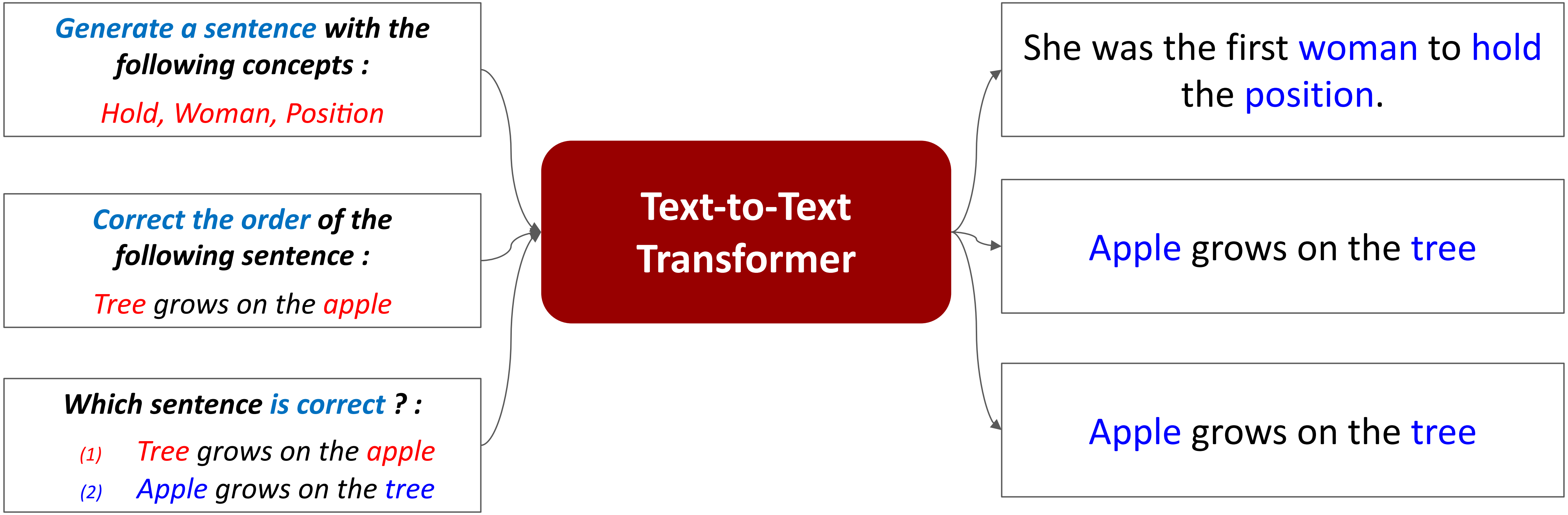
* equal contribution

USC University of Southern California

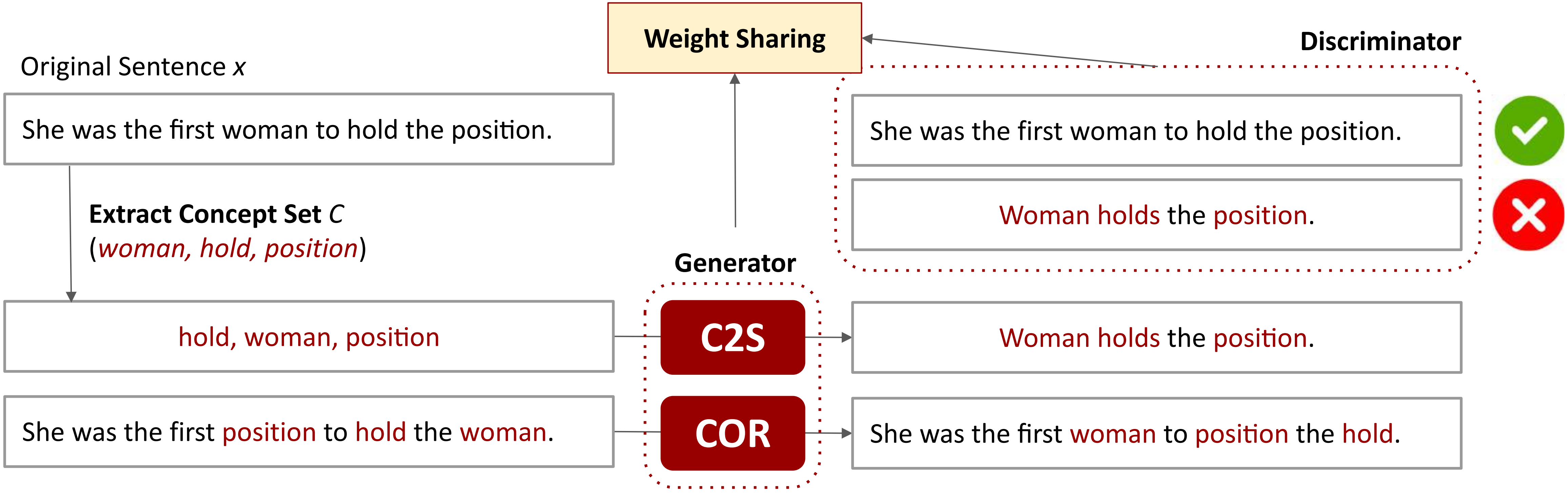
ISI

ink

Our idea : Novel Self-supervised Objectives to improve common sense reasoning ability.



CALM : Concept-Aware Language Model



(1) Given an input sentence x (*"She was the first woman to hold the position."*), extract concept-set C (*woman, hold, position*).
(2) Given x and C, produce corrupted source sentence x' either for **C2S** and **COR**
(3) The **<u>generator</u>** trained with **C2S** and **COR** recovers sentence x' to distractor x''
(4) The **<u>discriminator</u>** is trained to distinguish truth sentence from distractor x''

Is **CALM** reason with concepts ? ***Yes !***

Methods	CSQA	OBQA	PIQA	aNLI
	Accuracy			
T5-base	61.88(±0.08)	58.20(±1.0)	68.14(±0.73)	61.10(±0.38)
T5-base w/ additional epochs	61.92(±0.45)	58.10(±0.9)	68.19(±0.77)	61.15(±0.52)
T5-base + SSM	62.08(±0.41)	58.30(±0.8)	68.27(±0.71)	61.25(±0.51)
CALM (Generative-Only)	62.28(±0.36)	58.90(±0.4)	68.91(±0.88)	60.95(±0.46)
CALM (Contrastive-Only)	62.73(±0.41)	59.30(±0.3)	70.67(±0.98)	61.35(±0.06)
CALM (Mix-only)	63.02(±0.47)	60.40(±0.4)	70.07(±0.98)	62.79(±0.55)
CALM (w/o Mix warmup)	62.18(±0.48)	59.00(±0.5)	69.21(±0.57)	61.25(±0.55)
CALM	**63.32(±0.35)**	**60.90(±0.4)**	**71.01(±0.61)**	**63.20(±0.52)**

Experimental results on commonsense reasoning dataset.

Methods	CSQA	OBQA	PIQA	aNLI
	Accuracy (official dev)			
BERT-large	57.06(±0.12)	60.40(±0.6)	67.08(±0.61)	66.75(±0.61)
T5-large	69.81(±1.02)	61.40(±1.0)	72.19(±1.09)	75.54(±1.22)
CALM-large (Mix-only)	70.26(±0.23)	62.50(±1.0)	73.70(±1.09)	75.99(±1.26)
CALM-large	71.31(±0.04)	**66.00(±1.0)**	75.11(±1.65)	77.12(±0.34)

Effective in Large Models.

Smooth Communication Requires Commonsense

Text Message:

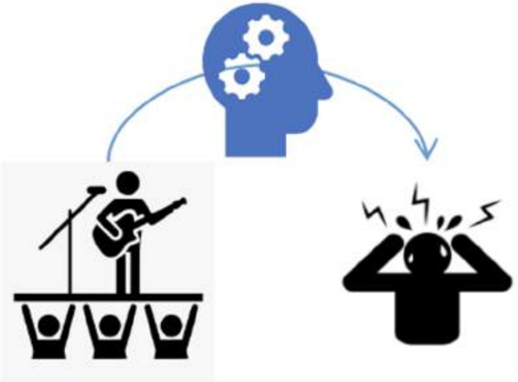
"I'm going to perform in front of thousands tomorrow…"

Explicit Knowledge:

Friend is going to perform in front of many people tomorrow

Commonsense Axiom:

Performing in front of people can cause anxiety



Text Message

"Deep breaths, you'll do great!"

Inference Made:

My friend might be anxious, let me try to calm them

Linguistically-Varied Statements of the same Commonsense Axiom

- A person performing in front of people might be nervous
- People performing in front of people find it harder to be relaxed
- It can be hard for someone to be calm when they're about to perform

Two key challenges

Inference making requires *implicit* commonsense reasoning

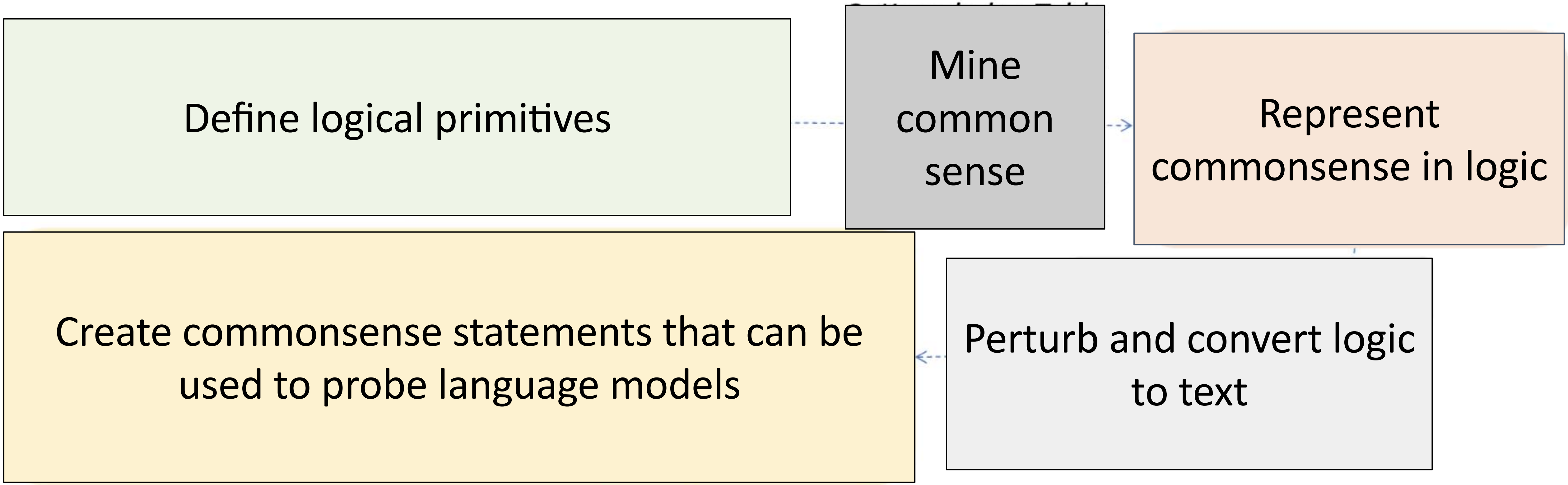
Humans fluidly adapt to *diverse* linguistic expressions

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-Findings 2020

The RICA Challenge

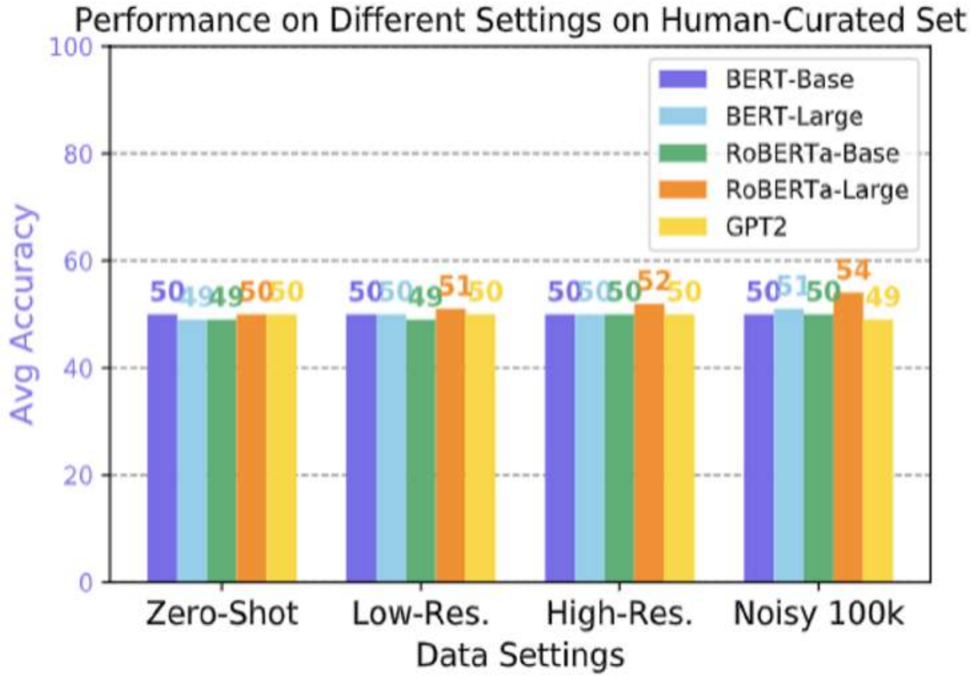


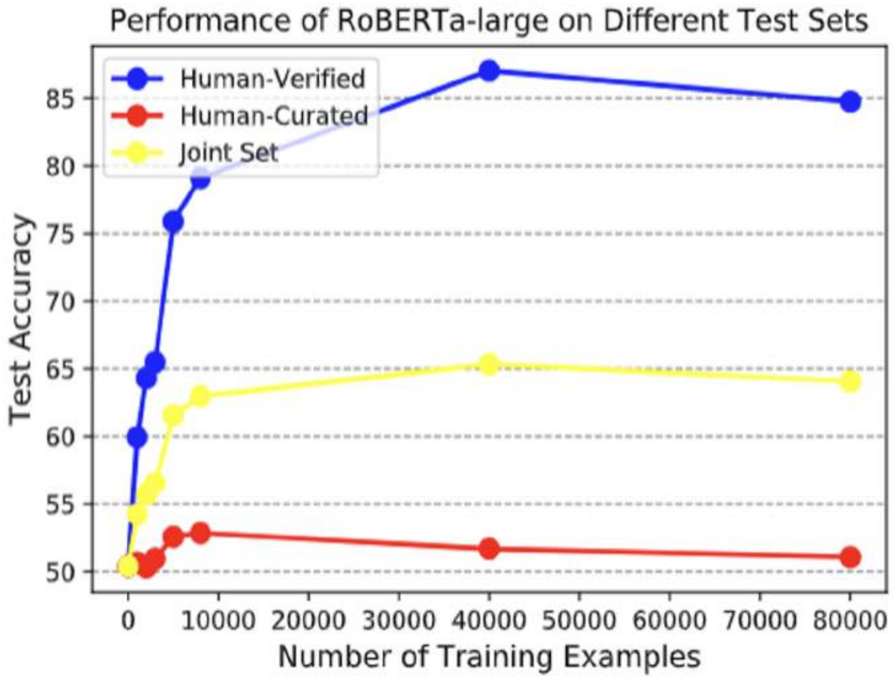
Results: **random guessing, heavy bias, and not robust**

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
- Curated-set provides great challenges for models

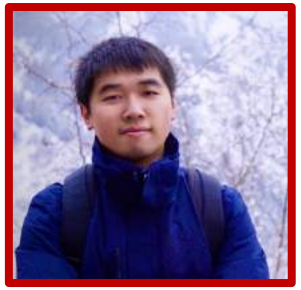
Human Performance: 91.7%





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

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.
- In this work, we refer to this ability as ***cross-task generalization***.
- We explore whether and 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

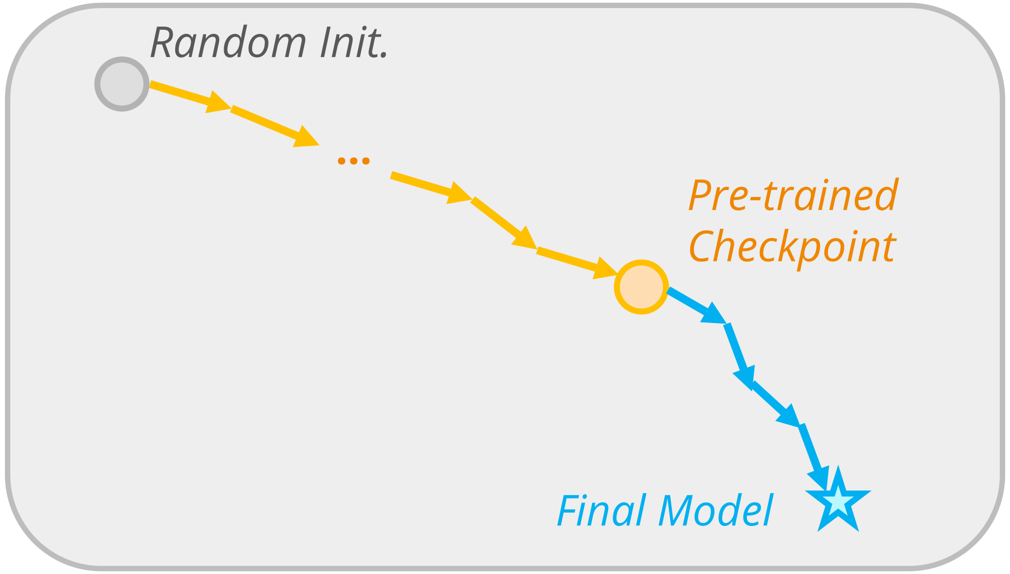
Learning Calculus in undergrad
Student: **Yes I can do it!**

Problem Setting

Prevalent Pipeline

Large-scale Pre-training

+ Downstream Fine-tuning

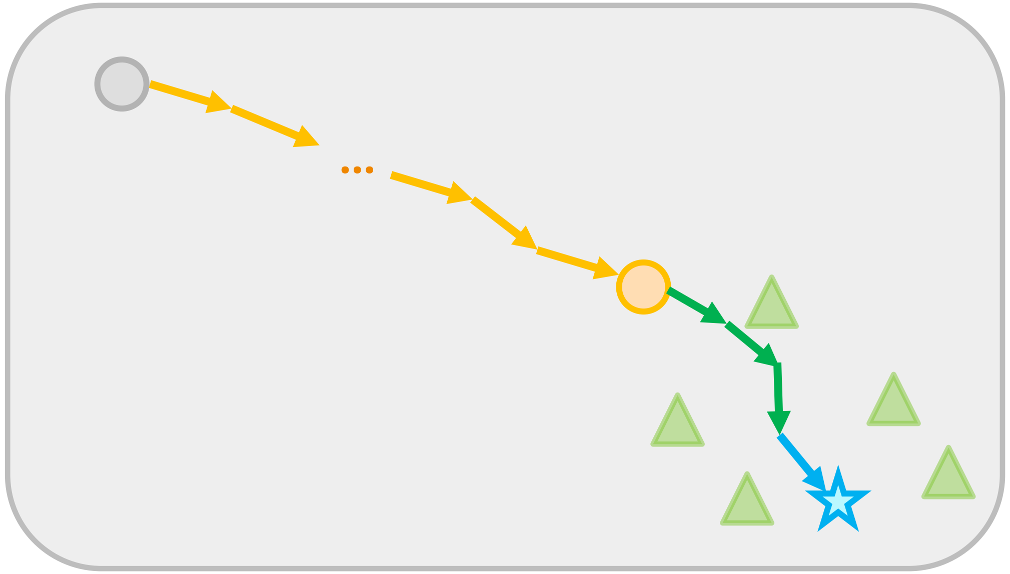


In our CrossFit 🏋️ Setting

Large-scale Pre-training

+ Upstream Learning on a set of seen tasks

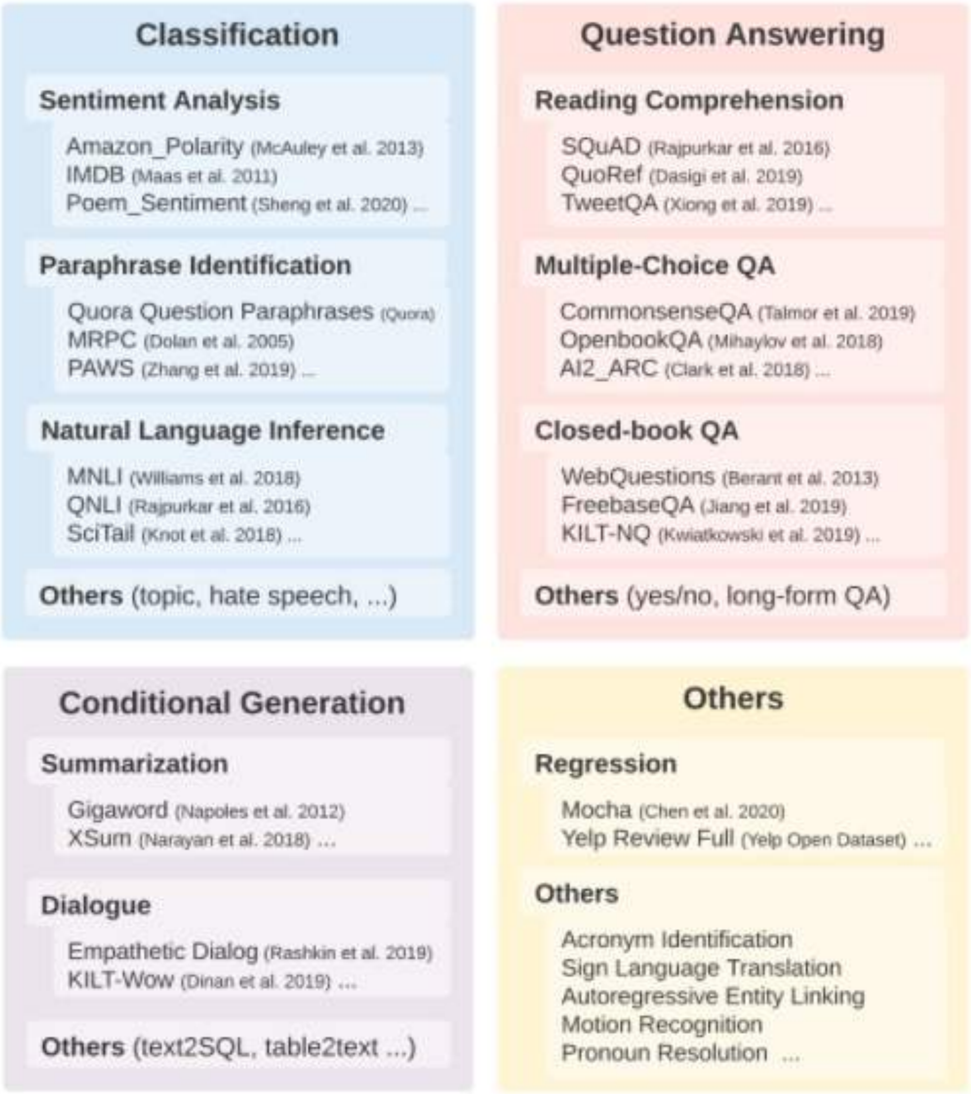
+ Downstream Fine-tuning on an unseen target task



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

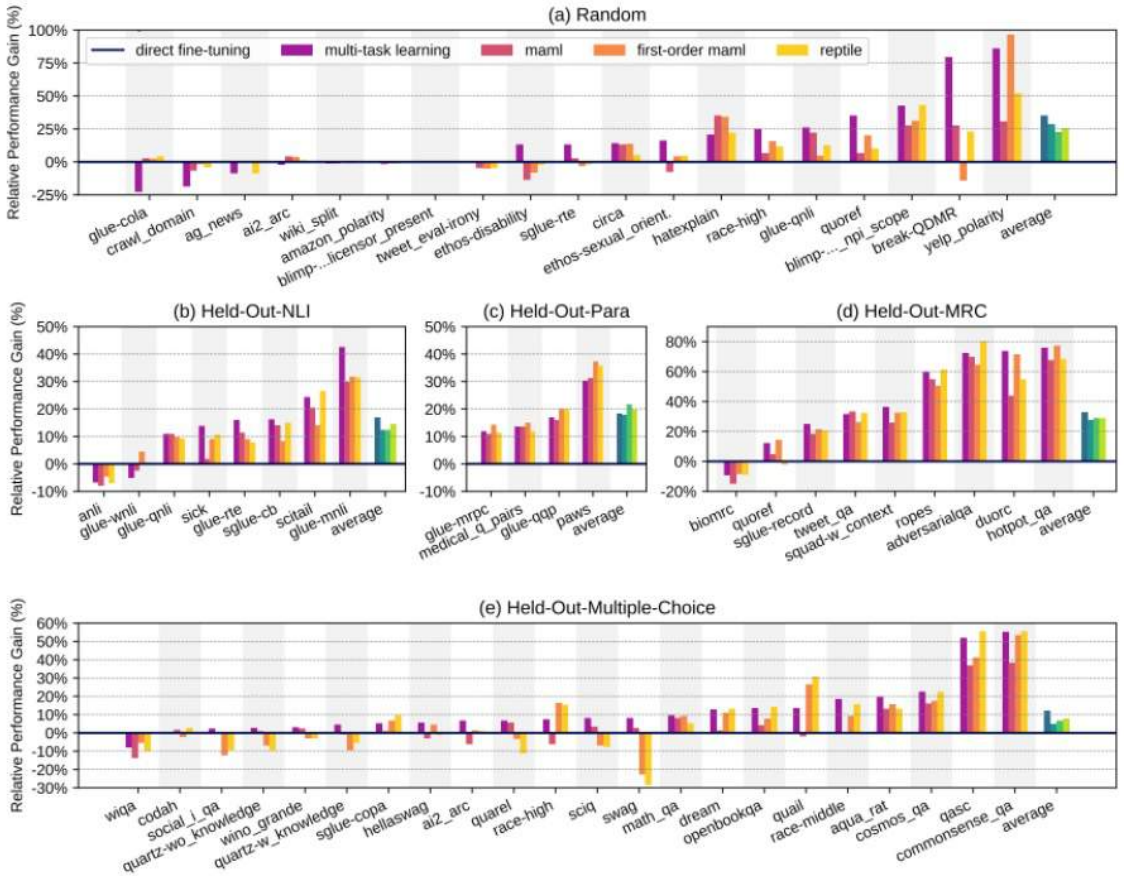


Key Findings

- **Q1. Can we teach pre-trained LMs to generalize across tasks with an upstream learning stage?**

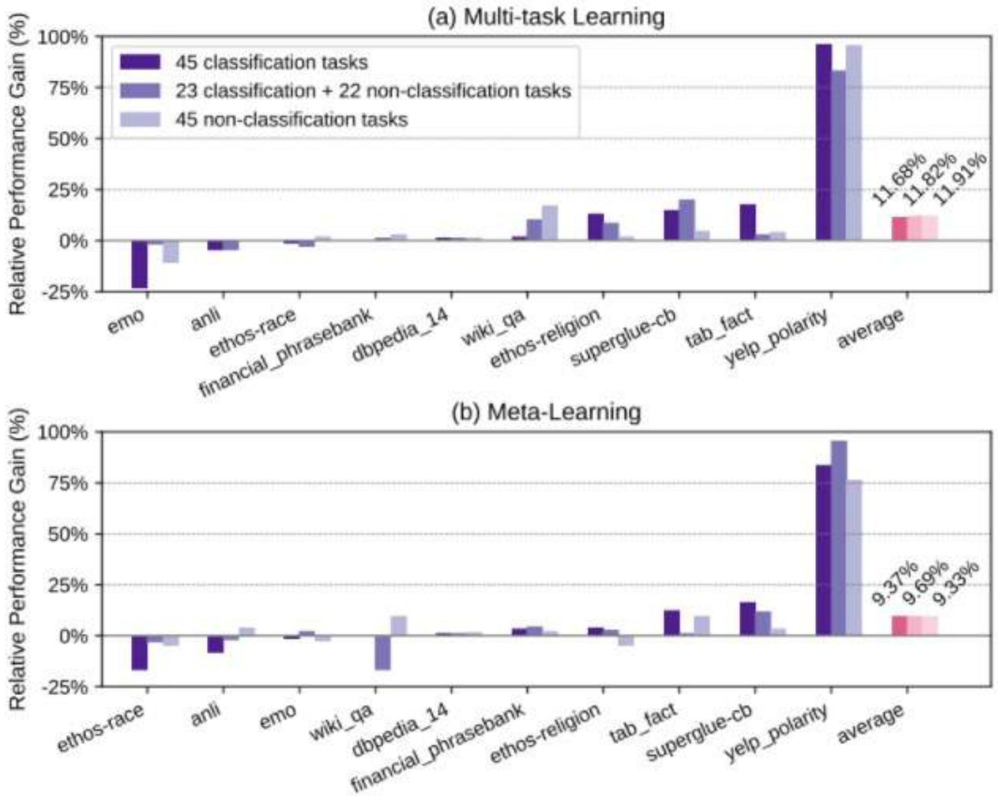
- We tried applying **multi-task learning** and **meta-learning** methods during the upstream learning stage.

- **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?**



- We conduct 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**.

Take-aways

- **CommonGen** is a task and dataset for generative commonsense reasoning in the format of NLG.
- **OpenCSR** is a challenge for open-ended CSR.
- CALM is a pre-trained language model for both discriminative and generative CSR tasks (including CommonGen).
- **RICA** analyzes and evaluates the robustness of NLU models based on logica and commonsense knowledge.
- **CrossFit** provides a standardized benchmark for developing and evaluating cross-task few-shot generalization.

→ Overall, we want to develop open-ended, robust, and generalizable AI systems with common-sense reasoning abilities.