Commonsense Reasoning in the Wild

Xiang Ren

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

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

Language is often ambiguous / underspecified

Hey, let's hoop at 10. Same park.

Q: Here, what does "10" mean?

Caption: Her voice is amazing!

Q: Who does "her" refer to?

Making proper presumptions is important!

Hey, let's hoop at 10. Same park.

Q: Here, what does "10" mean?

- *When meeting, people usually specify place and time*
- *Time can be referred by numbers*

A: 10 refers to time of day.

(Still not clear if it is 10AM or 10PM!)

Caption: Her voice is amazing!

Q: Who does "her" refer to?

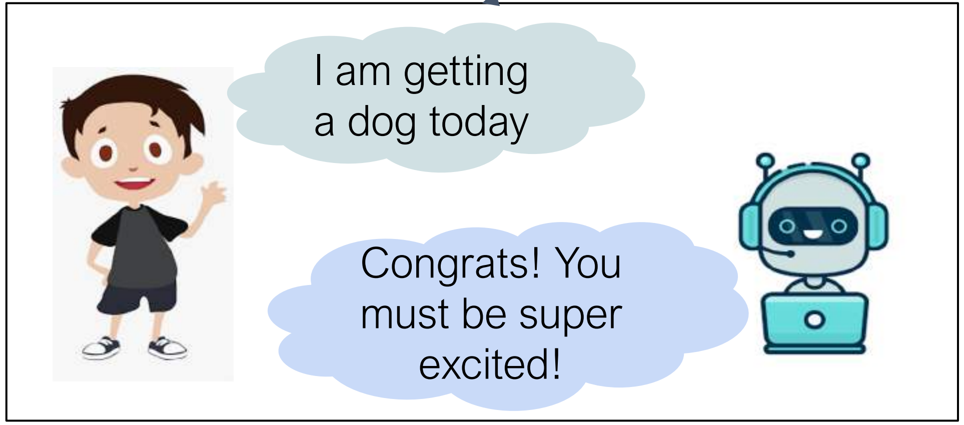
- *A person holding a microphone would have more prominent voice*
- *A person standing on a stage in front of an audience is likely singing/speaking*

A: "Her" refers to the girl in red dress.

Levinson 2020. Presumptive Meanings: The Theory of Generalized Conversational Implicature

Common sense knowledge are shared across tasks

A person feels happy and excited after getting a pet



Dialogue Response Generation

Q: How is the boy feeling right now?
A: stressed, happy, sad, confused

VQA

LM pretraining is not the answer

A bird usually has [MASK] legs.	1st: four (44.8%) 2nd: two (18.7%)
A car usually has [MASK] wheels.	1st: four (53.7%) 2nd: two (20.5%)
A car usually has [MASK] round wheels.	1st: two (37.1%) 2nd: four (20.2%)

Lin et al., 2020

Q: What color are the safety cones?
GT A: green
Predicted A: orange

Most cones were orange in training set

Agrawal et al., 2016

Premise: The judge by the actor stopped the banker.
Hypothesis: The banker stopped the actor.
Answer: Entailment

Lexical overlaps usually indicate entailment in training data

McCoy et al., 2019

LM pretraining is not the answer

A bird usually has [M

1st: four(44.8%)

A car usually has

A car usually has

Lin et al.,

lost cones were
range in training set

Reporting bias of commonsense knowledge <-> pretraining of massive language models

Premise:
the banke
Hypothesis:
Answer: Entailment ✗

Lexical overlaps usually indicate entailment in training data

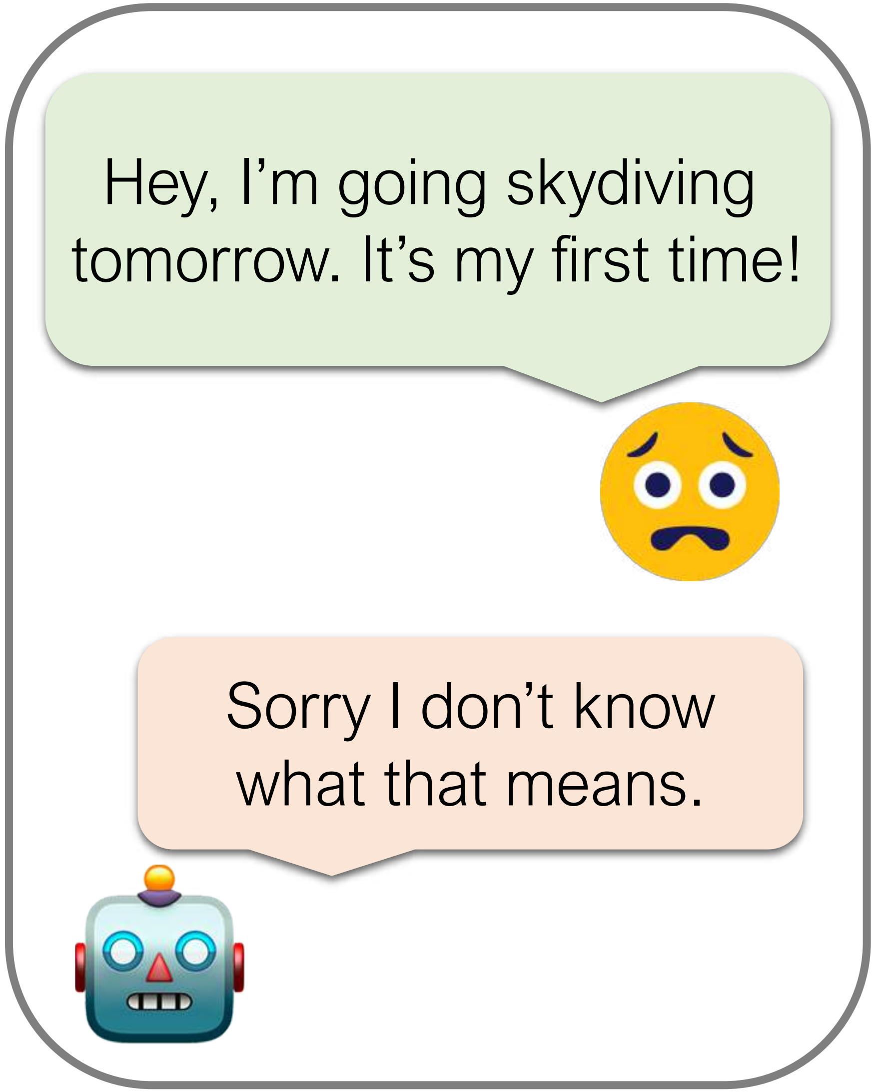
McCoy et al., 2019

CSR Models on
Research Benchmarks

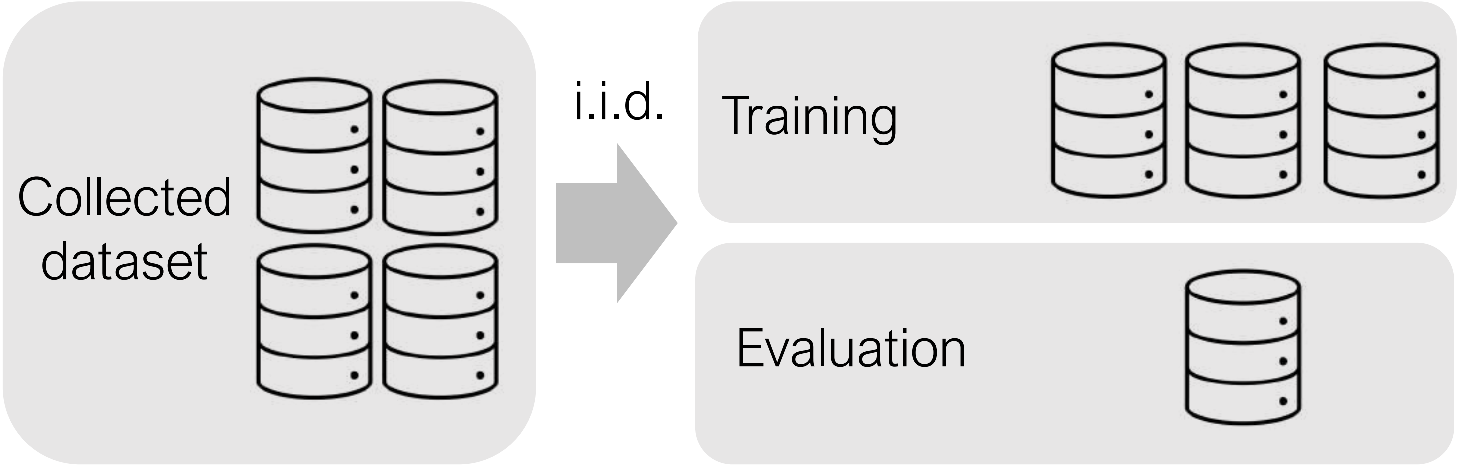
XYZ Leaderboard

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

CSR Models in the Wild

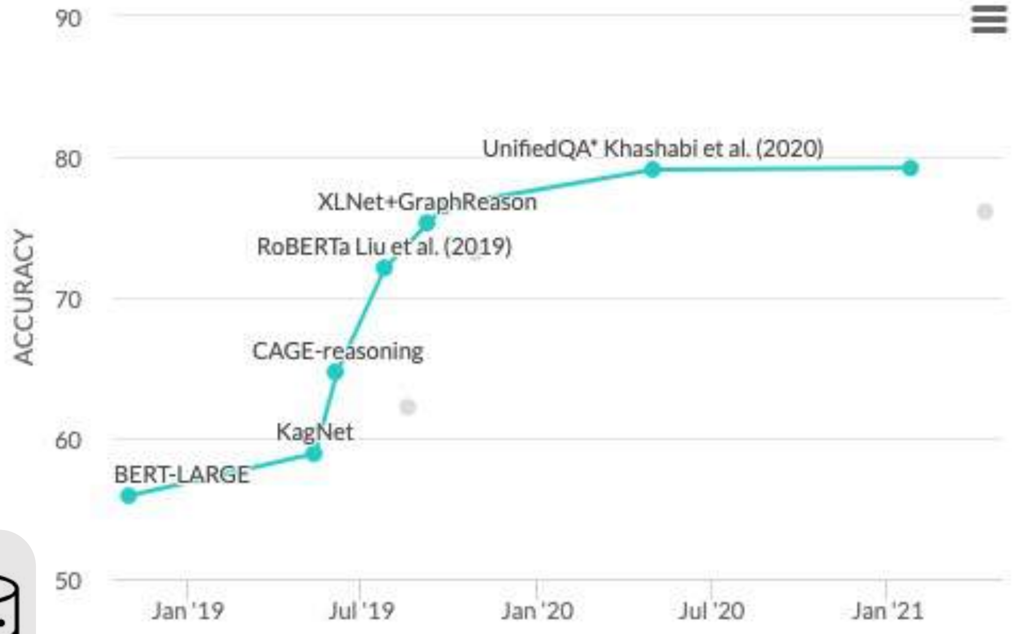


Performs well on a benchmark



97% Acc.

Commonsense Question Answering



Paper With Code: CommonsenseQA 1.1

Performs well **on a benchmark**

- Model learns dataset shortcuts

Performs well **in the wild**

- Robust to linguistic variations

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

Linguistically-varied statements of the same inference rule

RICA (Zhou et al., 2021)

Behavioral Testing

Ribeiro et al., 2020

Robust.	***INV:*** Swap one character with its neighbor (typo)
	DIR: Paraphrase of question should be duplicate

Performs well **on a benchmark**

- Model learns dataset shortcuts
- **Struggles with underspecified/adversarial inputs**

Performs well **in the wild**

- Robust to linguistic variations
- **Resolves ambiguity/noise with presumptions**

When is the Super Bowl?

Search

Do you mean When is the Super Bowl 2022?

Super Bowl 2022 will be at 3:30 PM on February 13.

Underspecified Inputs

Levinson, 2000

Adversarial Inputs

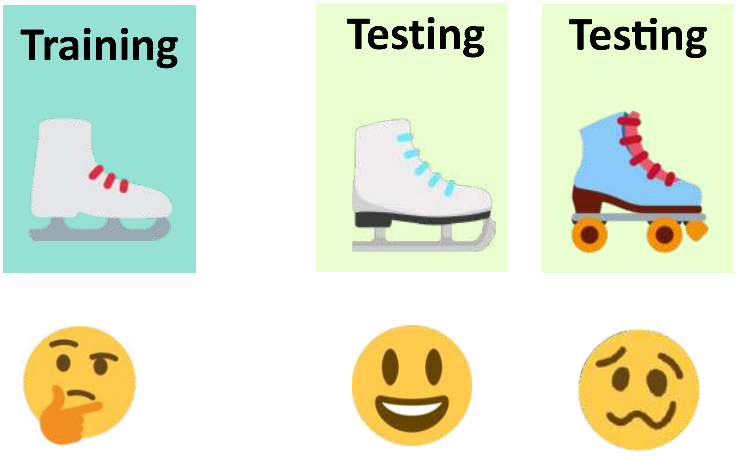
Jia and Liang, 2017
Wallace et al., 2019

Performs well on a benchmark

- Model learns dataset shortcuts
- Struggles with underspecified/adversarial inputs
- Customized to a narrow task

Performs well in the wild

- Robust to linguistic variations
- Resolves ambiguity/noise with presumptions



Performs well **on a benchmark**

- Model learns dataset shortcuts
- Struggles with underspecified/adversarial inputs
- **Customized to a narrow task**

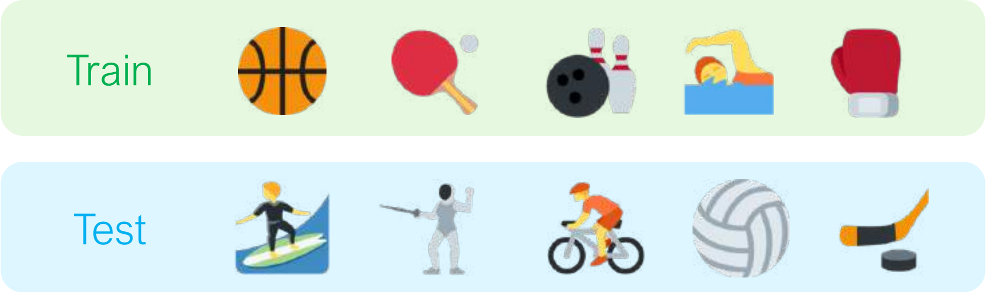
Performs well **in the wild**

- Robust to linguistic variations
- Resolves ambiguity/noise with presumptions
- **Generalizable across a wide range of tasks**

applicable to a wide range of tasks

DecaNLP (McCann et al., 2018)
T5 (Raffel et al., 2019)
ExT5 (Aribandi et al., 2021)
Muppet (Aghajanyan et al., 2021)

generalizes well to new tasks



CrossFit (Ye et al., 2021)
Natural Instructions (Mishra et al., 2021)
FLEX (Bragg et al., 2021)
FLAN (Wei et al., 2021)
T0 (Sanh et al., 2021)

This talk - New ways of formulating CSR challenges

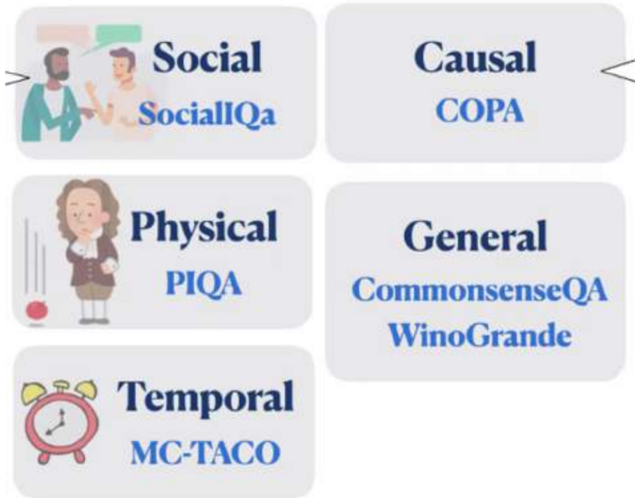
Discriminative (closed-ended) reasoning

Alex spilled the food she just prepared all over the floor and it made a huge mess.

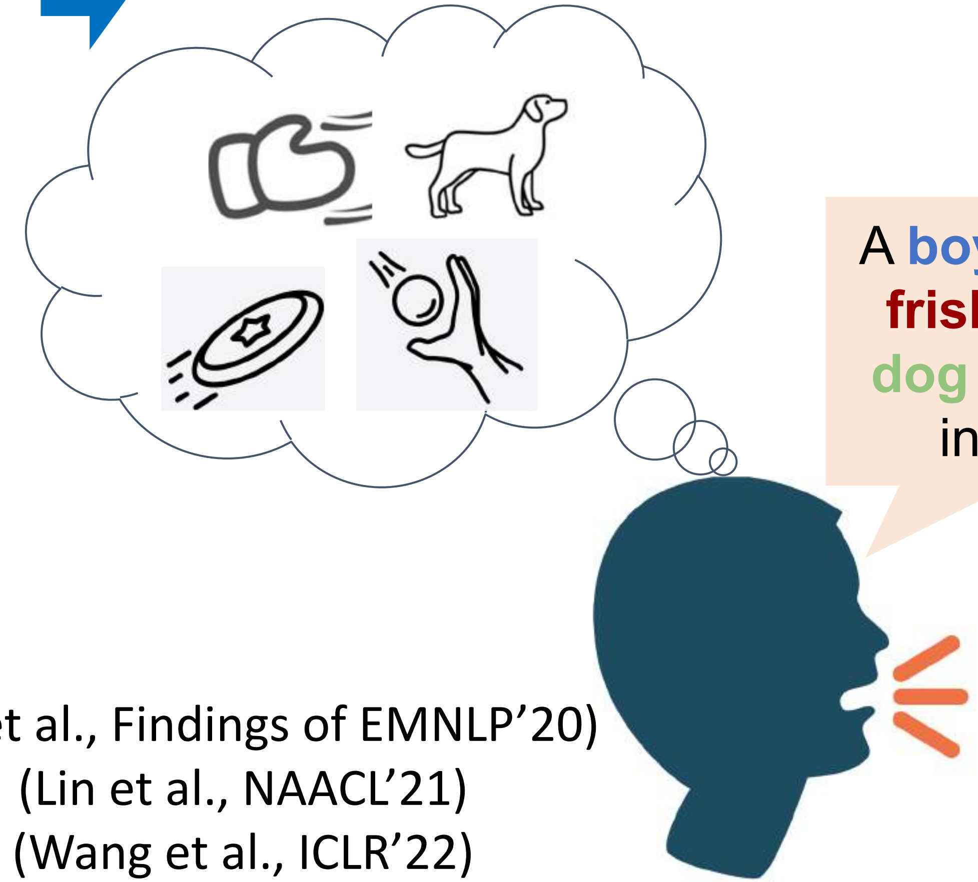
Q What will Alex want to do next?

A (a) taste the food
(b) mop up ✓
(c) run around in the mess

Social IQA (Sap et al. 2019)



Towards more **open-ended** reasoning



A **boy throws** a **frisbee** and a **dog catches** it in the air.

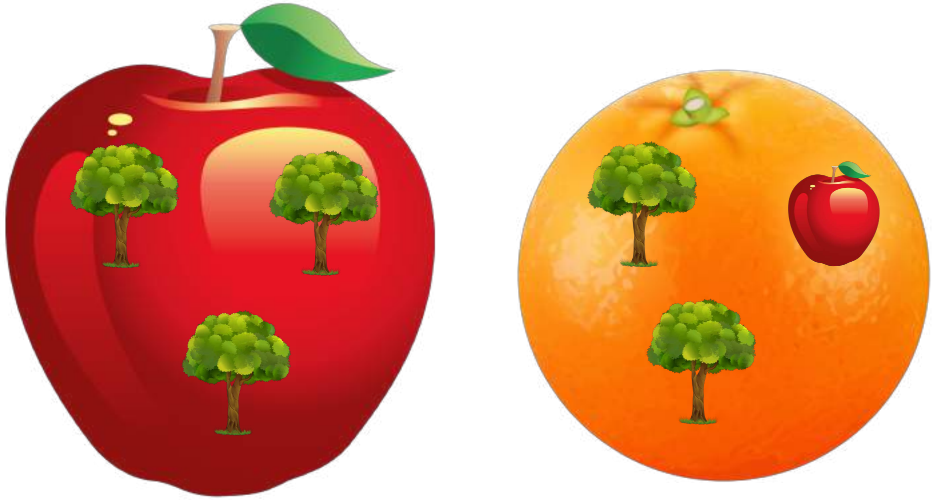
(Lin et al., Findings of EMNLP'20)
(Lin et al., NAACL'21)
(Wang et al., ICLR'22)

This talk - New ways of formulating CSR challenges

Reasoning in a logically robust/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 and oranges~~
~~Apples and trees grow on oranges~~



(Zhou et al., EMNLP'21)

This talk - New ways of formulating CSR challenges

Study the cross-task generalization ability of NLP models

Directly Learning Calculus…

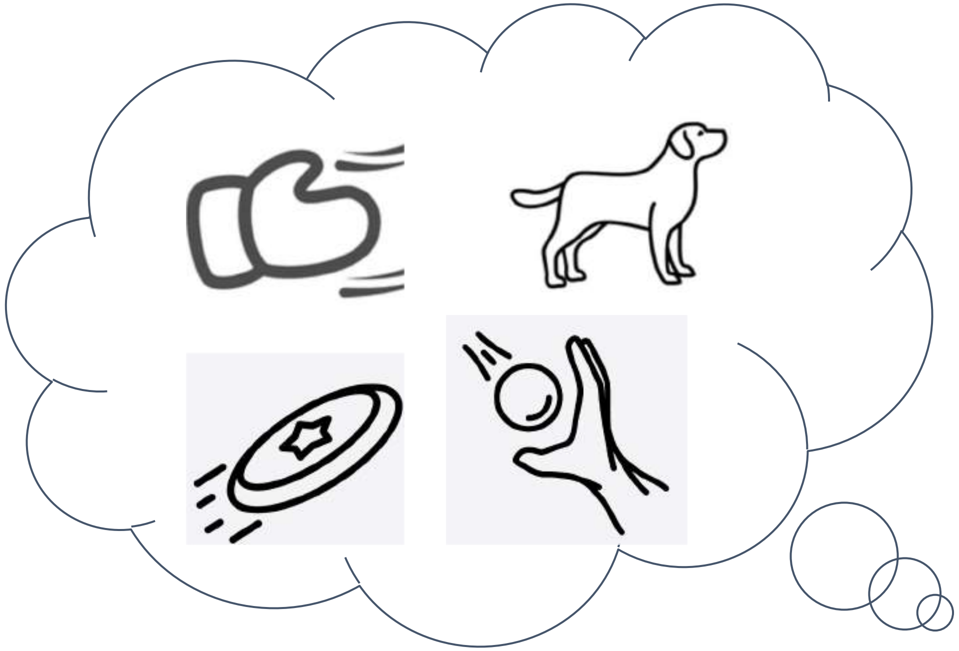
Machine: ? ? ?
Baby: ? ? ?

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

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

(Ye et al., EMNLP’21)

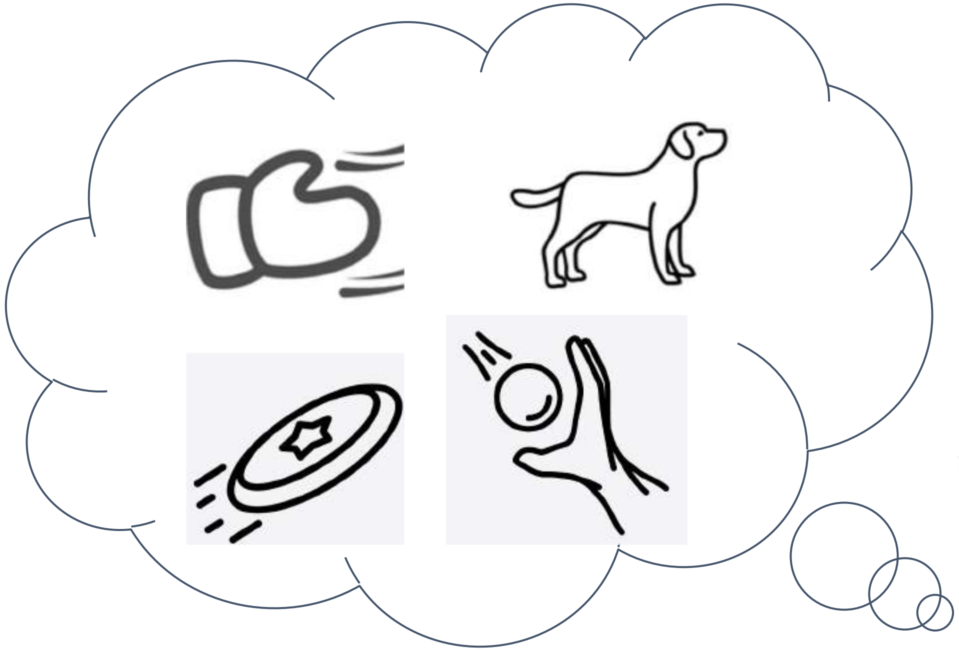
An intelligent behavior possessed by humans that demonstrates common sense



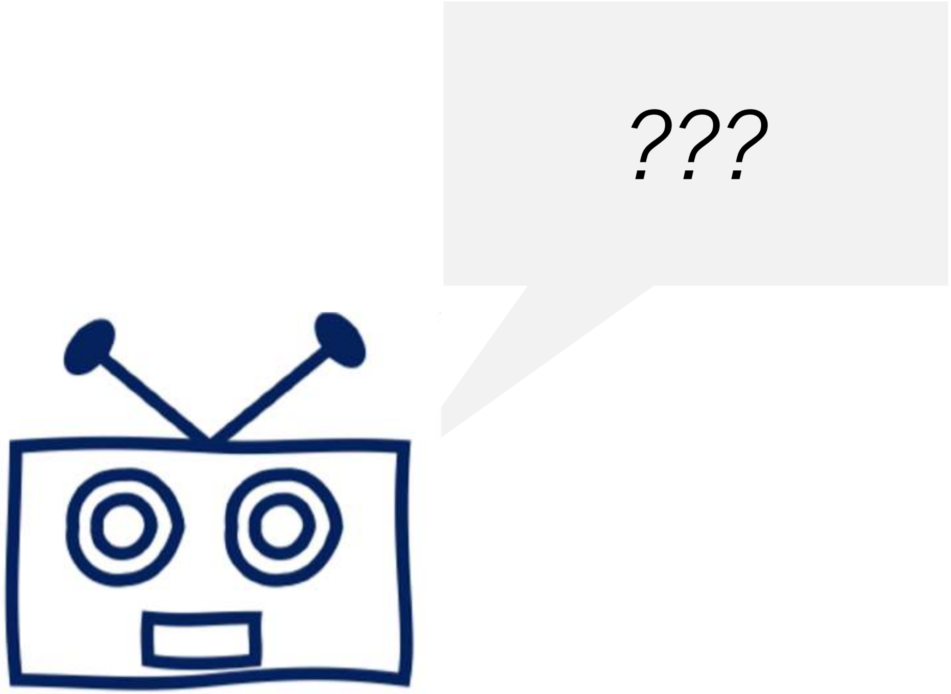
{dog, frisbee, catch, throw}

A boy throws a frisbee and a dog catches it in the air.

Can machines learn to describe a daily scene using concepts?



{dog, frisbee, catch, throw}



Generative Commonsense Reasoning

Input: A set of concept words (objects / actions)

Output: A sentence describing everyday scenes using all the concepts.

Statistics	Train	Dev	Test
# Concept-Sets	**32,651**	**993**	**1,497**
-Size = 3	25,020	493	-
-Size = 4	4,240	250	747
-Size = 5	3,391	250	750
% Unseen Concepts	-	6.53%	8.97%
% Unseen Concept-Pairs	-	96.31%	100.00%
% Unseen Concept-Triples	-	99.60%	100.00%

({dog, frisbee, catch, throw})

Humans

A dog catches a frisbee when a boy throws it.

Machines

GPT2: A dog throws a frisbee at a football player.

T5: Dog catches a frisbee and throws it at a dog.

(**CommonGen**, Findings of EMNLP 2020)

Bill Yuchen Lin[♥] Wangchunshu Zhou[♥] Ming Shen[♥] Pei Zhou[♥]
Chandra Bhagavatula[♠] Yejin Choi[♠♣] Xiang Ren[♥]

USC University of Southern California
UNIVERSITY of WASHINGTON
AI2

Rank	Model	BLEU-4	CIDEr	SPICE
	Upper Bound	46.49	37.64	52.43
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** *Alibaba and Xiamen University.* Email Paper (AAAI 2022)	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
6 Jan 28, 2022	**Imagine-and-Verbalize** *USC/ISI* Email Paper (ICLR22)	40.565	17.716	31.291
7 Jan 13, 2021	**RE-T5 (Retrieval-Enhanced T5)** *Microsoft Cognitive Services Research Group* Email Paper (ACL21)	40.863	17.663	31.079
8 Oct 19, 2021	**A* Neurologic (T5-large)** *UW and AI2* Email Description	39.597	17.285	30.130
9 Aug 1, 2021	**VisCTG (BART-large)** *CMU-LTI* Email Paper (arXiv)	36.939	17.199	29.973

Datasets

GEM

(Gehrmann et al., 2021)

T0

(Sanh et al., 2021)

FLAN

(Wei et al., 2021)

CommonGen Leaderboard: https://inklab.usc.edu/CommonGen/leaderboard.html

Rank	Model	BLEU-4	CIDEr	SPICE
	Upper Bound	46.49	37.64	52.43
Aug 1, 2021	CMU-LTI Email Paper (arXiv)	36.939	17.199	29.973

Datasets

[Machine generations] {cow, horse, lasso, ride}

[bRNN-CpNet]: Someone lowers his horse from the wall and lasso glass by cows.

[Trans-CpNet]: A horse having lasso in the bridal cows.

[MP-CpNet]: Cow in a lasso getting the ride.

[LevenTrans]: A cow rides through a horse.

[GPT-2]: A horse rides on a lasso.

[BERT-Gen]: A cow rides a lasso on a horse.

[UniLM]: A man rides a horse with a lasso at cows.

[UniLM-v2]: A horse rides a cow with a lasso on it.

[BART]: A man rides a horse and a cow on a bridle with a lasso.

[T5]: Lasso to ride a cow on a horse.

GEM

ann et al., 2021)

T0

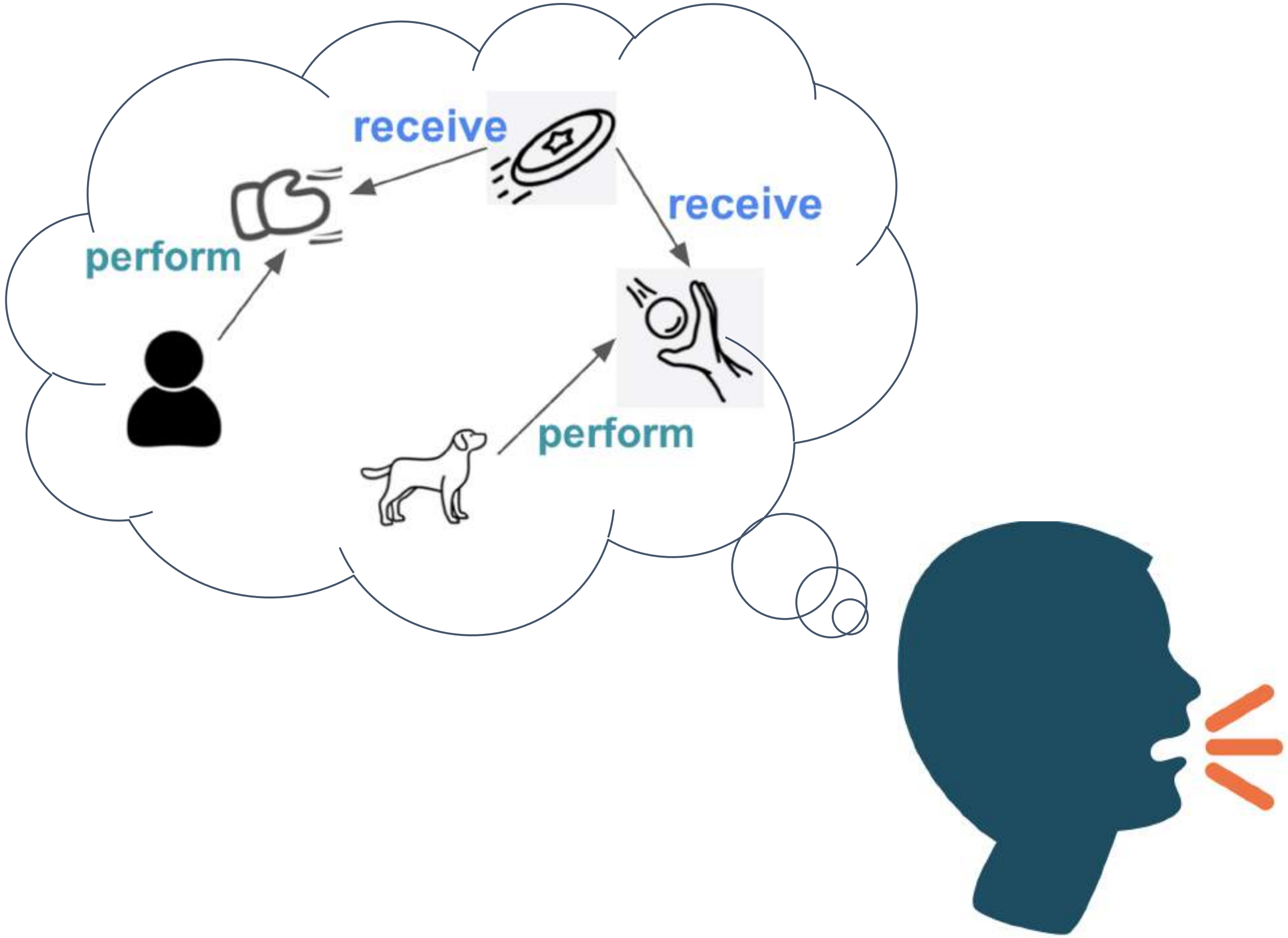
h et al., 2021)

FLAN

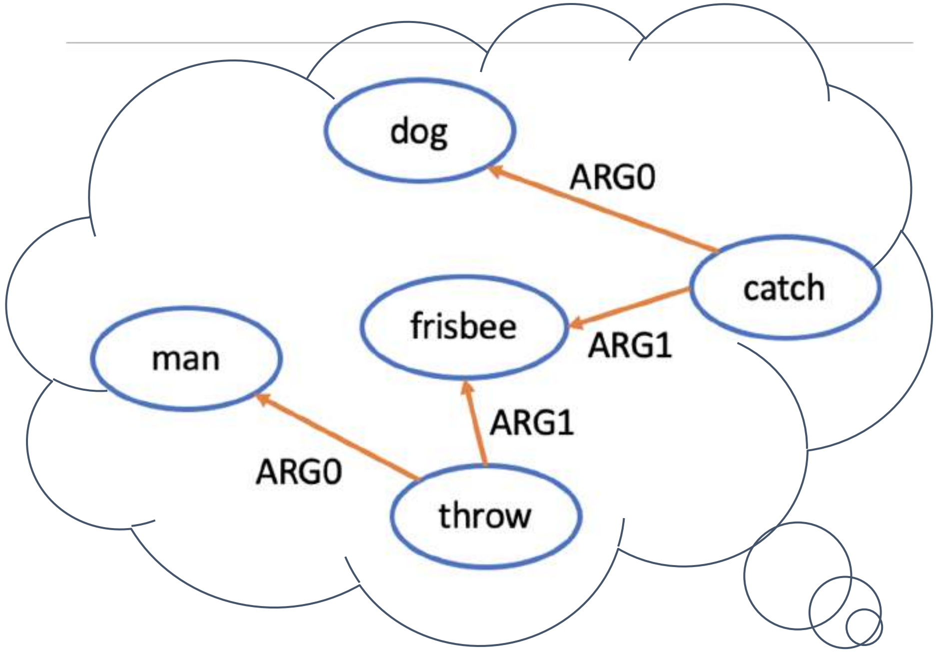
(Wei et al., 2021)

CommonGen Leaderboard: https://inklab.usc.edu/CommonGen/leaderboard.html

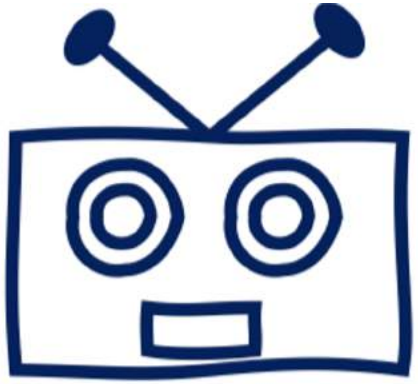
Externalizing scene imagination: Structured Knowledge Representation



Externalizing scene imagination: Structured Knowledge Representation

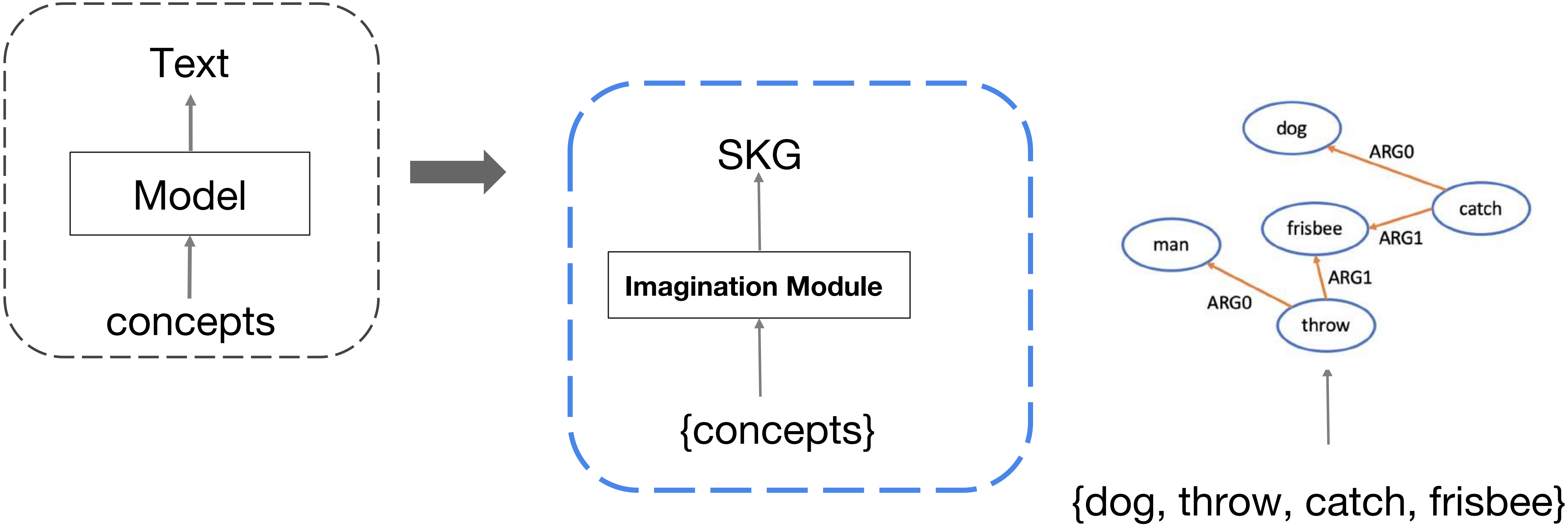


Scene Knowledge Graph (SKG)



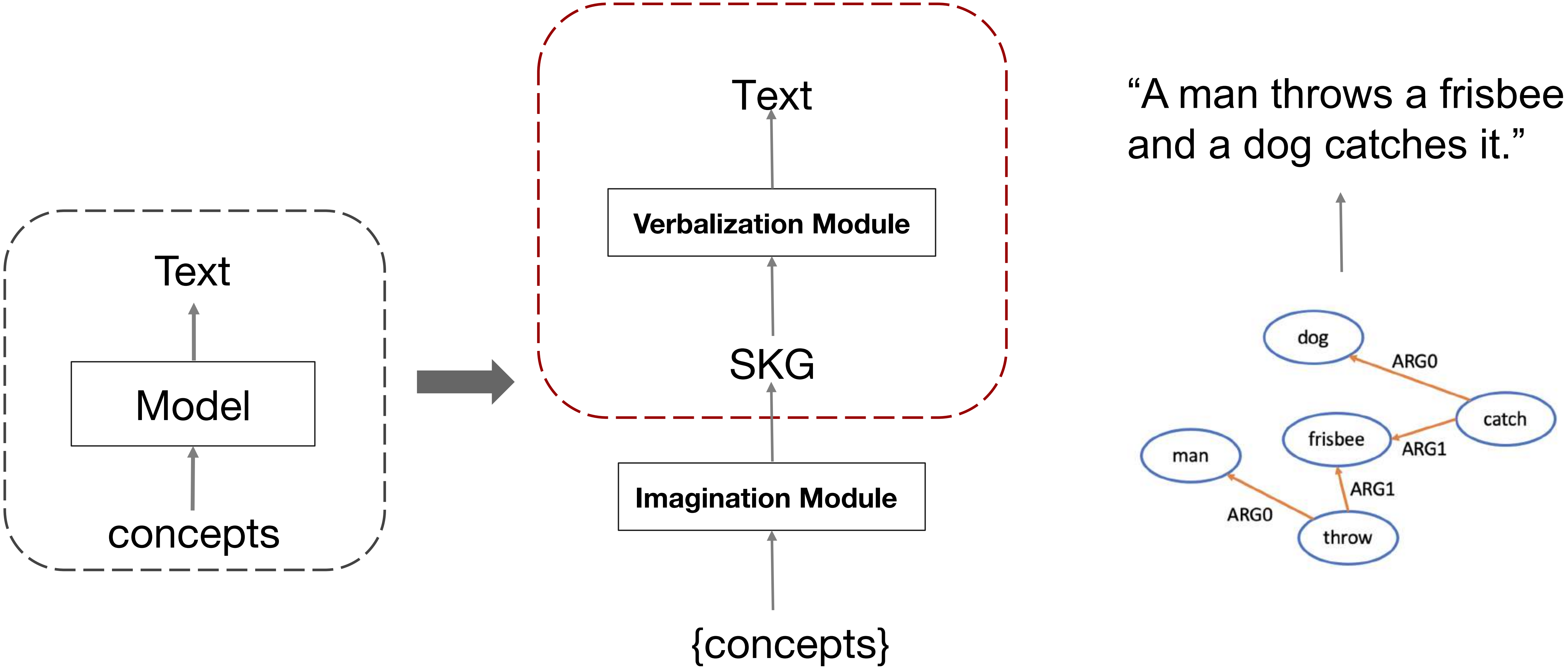
Relation types	Examples
ARG1	(play, ARG1, guitar)
ARG0	(play, ARG0, man)
ARG2	(ask, ARG2, girl)
Location	(play, Location, stage)
Time	(play, Time, sing)
Op1	(down, Op1, stair)
Part	(dog, Part, ear)

Externalizing scene imagination: **Imagine-and-verbalize**



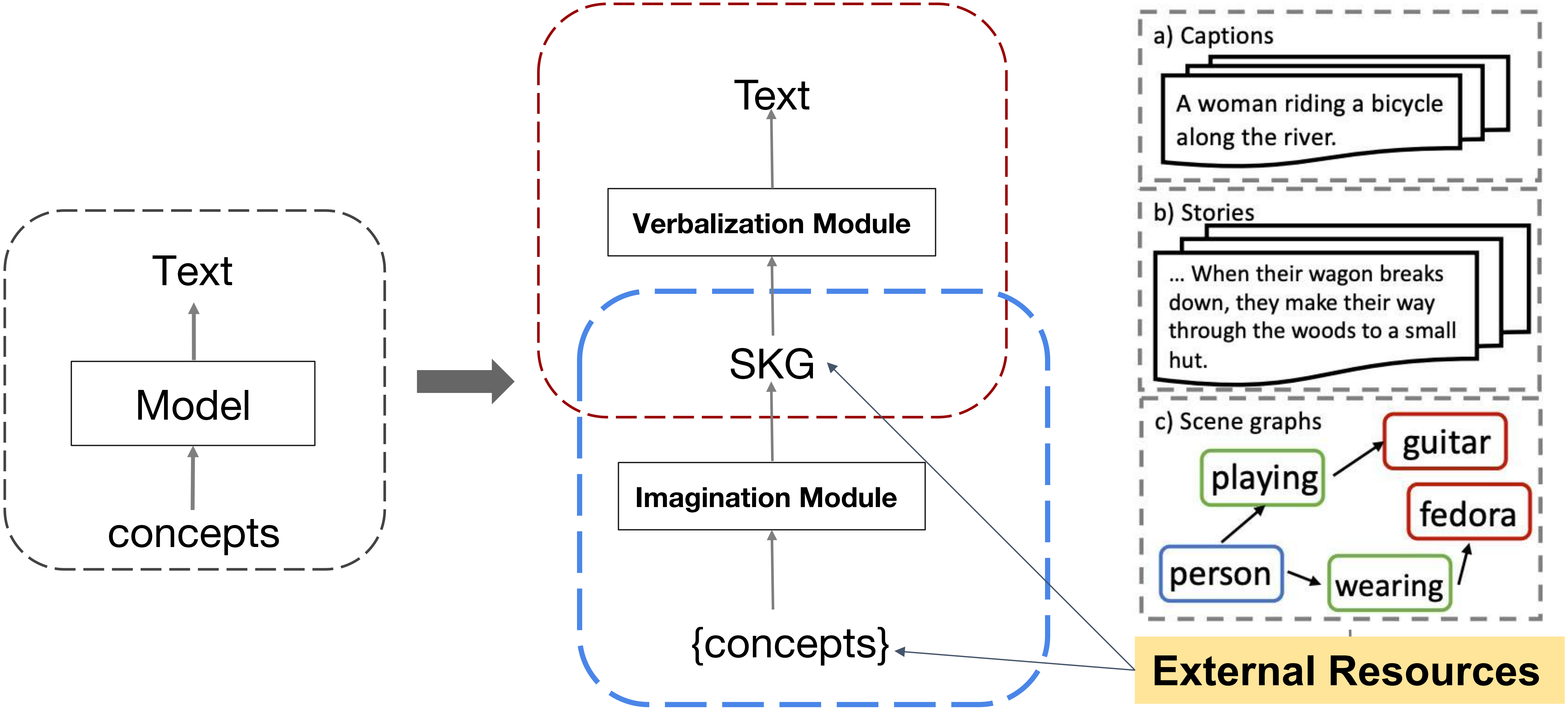
(Wang et al., ICLR'22)

Externalizing scene imagination: **Imagine-and-verbalize**



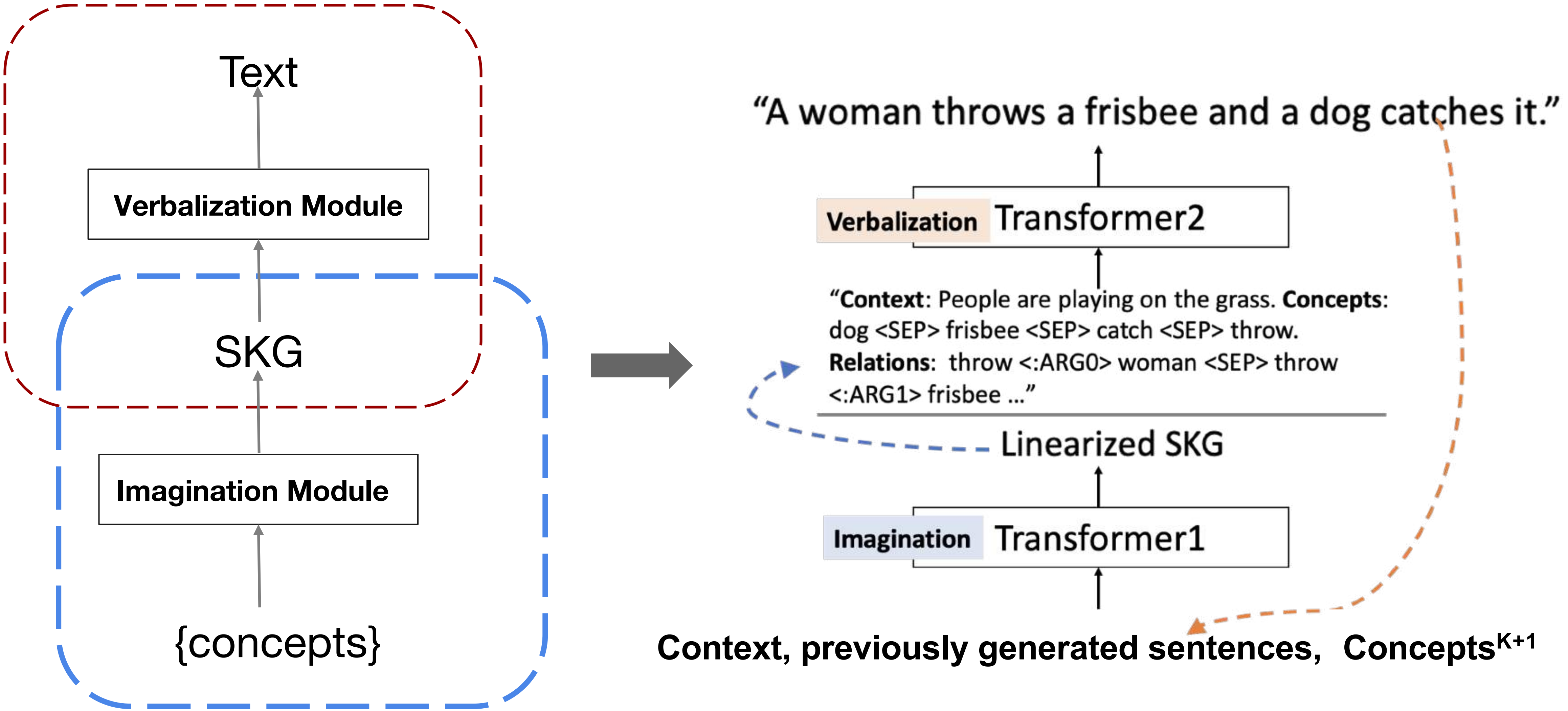
(Wang et al., ICLR’22)

Externalizing scene imagination: **Imagine-and-verbalize**



(Wang et al., ICLR'22)

Externalizing scene imagination: **Imagine-and-verbalize**



(Wang et al., ICLR’22)

How can imagination help?

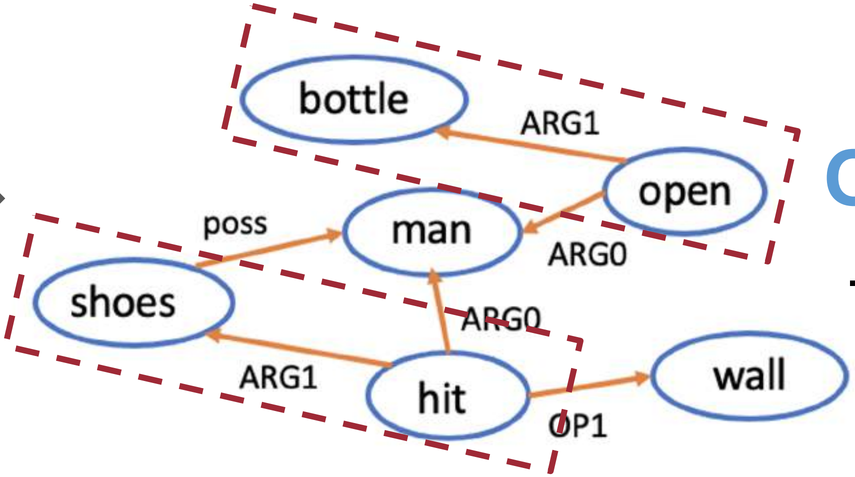
Input: {hit, bottle, open, shoe, wall}

without imagination

Output:

- Someone **opens** his **shoes** and hits a bottle on a wall.

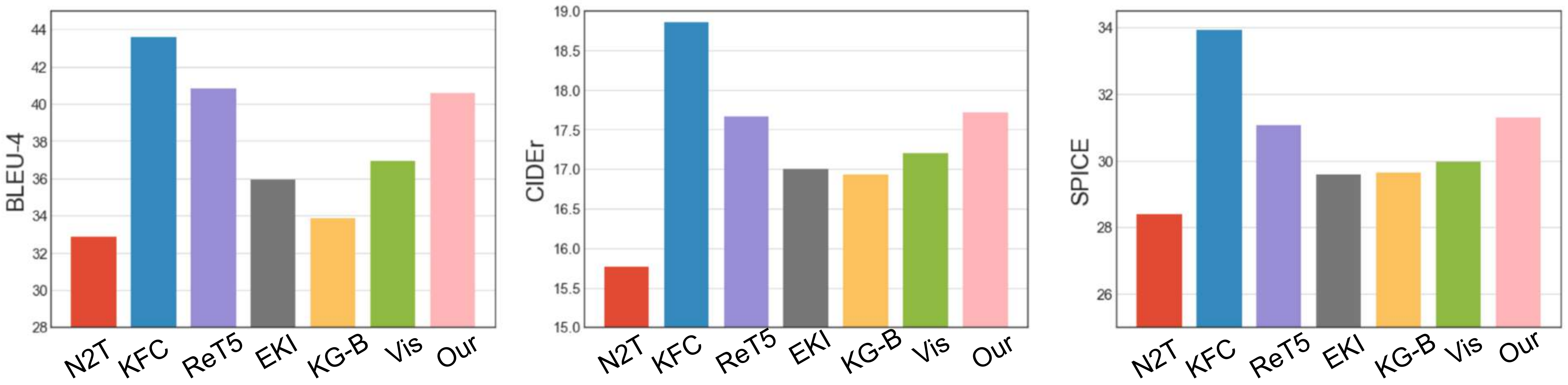
with imagination



Output:

- A man opens a bottle and hits his shoes against a wall.

Results on CommonGen (leaderboard)



- SOTA (KFCNet) uses a much larger corpus (>700M)
- Imagination > Prototype-based (Except KFCNet)
  > VisCTG (Image)
  > KG-BART
  > Node2Text

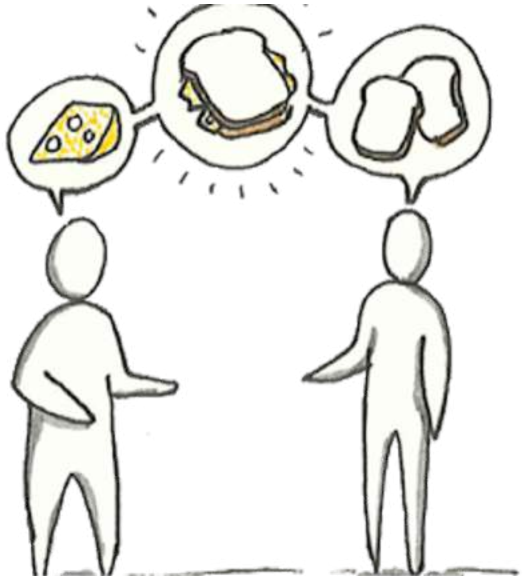
How do we reply in conversations?

I’m going to sing in front of hundreds tomorrow...

Performing in front of audience can cause ***anxiety***

Deep breaths, you’ll do great!

Grounding in Communications



Effective communications require reaching mutual beliefs and knowledge among participants (called *grounding*)

Common Sense plays a critical role in grounding in communications

Clark, H. H., & Brennan, S. E. (1991). *Grounding in communication.*

How do we reply in conversations?

I'm going to perform in a piano recital tomorrow...

Performing in front of audience can cause ***anxiety***

Deep breaths, you'll do great!

Recalling & Combining common sense with information expressed in NL to ***make inferences***

How do we reply in conversations?

I'm going to perform in a piano recital tomorrow...

Performing in front of audience can cause ***anxiety***

Deep breaths, you'll do great!

Recalling & Combining common sense with information expressed in NL to ***make inferences***

Producing ***consistent*** inferences amidst ***logically-equivalent yet linguistically-varied*** paraphrases

RICA: Robust Inference on Commonsense Axioms

- Sets of natural language statements in the ***“premise-conclusion"* format** that **express the *same commonsense axiom* but *linguistically varied***

- Examples:
  - **Original**: “A is heavier than B, so A is <better> at sinking than B.”
  - **Negation**: “A is heavier than B, so A is **not** <worse> at sinking than B.”
  - **Entity Swap**: “**B** is heavier than **A**, so A is <worse> at sinking than B.”
  - **Antonym**: “A is heavier than B, so A is <worse> at **floating** than B.”
  - …

Recalling & Combining common sense with information expressed in NL to ***make inferences***

(Zhou et al., EMNLP’21)

RICA: Robust Inference on Commonsense Axioms

- Probe model's ***robustness against linguistic variations*** (of the same commonsense axiom)

- Masked word prediction task: **Choose <better> or <worse>**:
  - **Original**: "A is heavier than B, so A is <MASK> at sinking than B."
  - **Perturb1**: "A is heavier than B, so A is **not** <MASK> at sinking than B."
  - **Perturb2**: "**B** is heavier than **A**, so A is <MASK> at sinking than B."
  - **Perturb3**: "A is heavier than B, so A is <MASK> at **floating** than B."
  - …

Producing ***consistent*** inferences amidst ***logically-equivalent yet linguistically-varied*** paraphrases

(Zhou et al., EMNLP'21)

RICA: Overview of the probe construction

Define logical primitives

Mine common sense

Represent commonsense in logical form

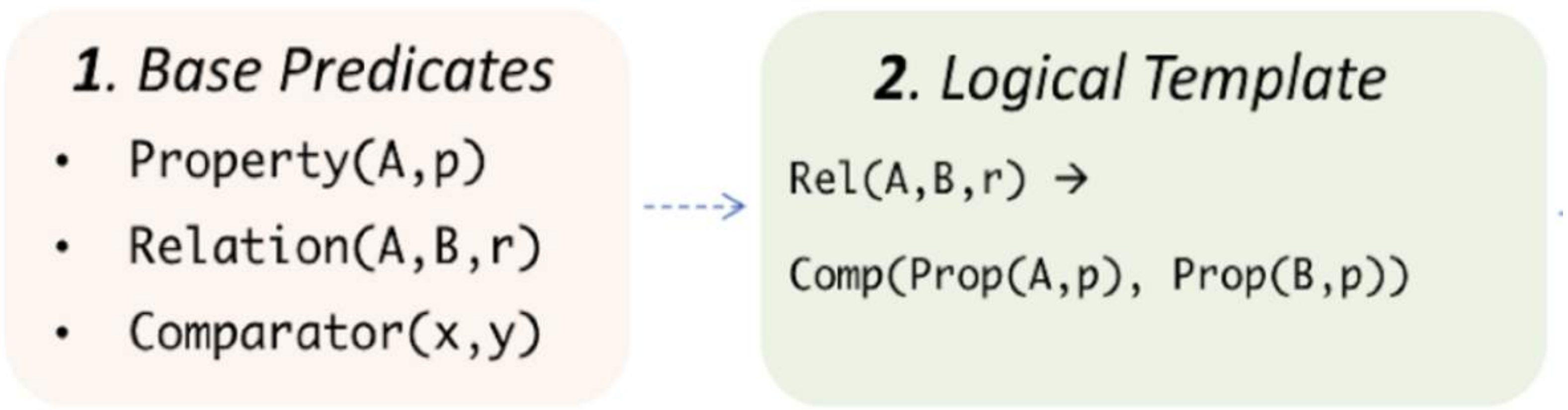
Perturb and convert logical form to text

Create commonsense statements that can be used to probe language models

(Zhou et al., EMNLP’21)

Probe construction I

Define logical primitives



- Define three basic first-order logic predicates
- Connect predicates to form abstract logical templates
    - *A is B 's <r>, so A is more/ less <p> than B*

Probe construction II

- **Goal:** Fill the abstract templates with concrete common sense

A is B's <r>, so A is more/less <p> than B
  - <r> → "*lawyer*"
  - <p> → "*knowledge of law*"

- Crawl from knowledge bases
  - Step 1: Get a list of occupations
  - Step 2: Query ConceptNet for triples, such as <Occupation, HasProperty, p>

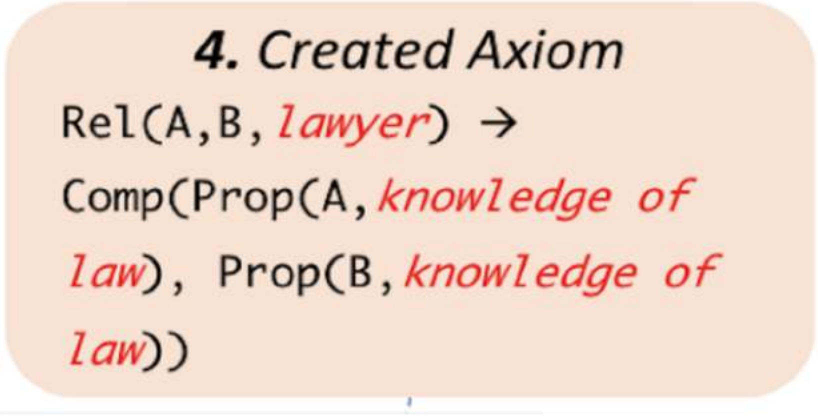
Mine common sense

3. Knowledge Table

Relation	Property
Lawyer	Knowledge of Law
Doctor	Takes care of people
...	...

Probe construction III

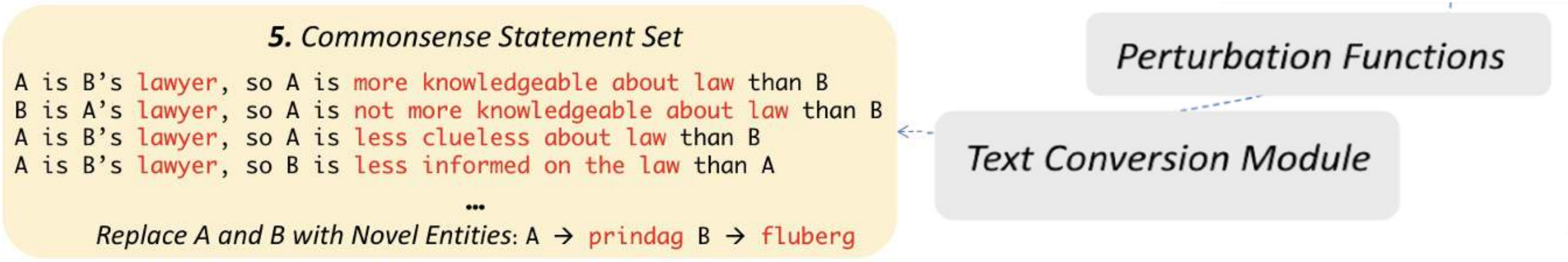
- Fill logical templates with crawled common sense



Represent commonsense in logical form

- Apply perturbation operators and convert to text

Perturb and convert logical form to text



Probe construction III

Perturb and convert logical form to text

Perturbation Functions

Goal: create perturbed forms that preserve the commonsense axiom

- **Linguistic Operators:**
  - Negation: "*knowledgeable*" → "*not knowledgeable*"
  - Antonym: "*knowledgeable*" → "*clueless*"
  - Paraphrase: "*knowledgeable*" → "*informed*"
  - **Composition:**
    - negation + paraphrase → "*not informed*"
    - …
- **Asymmetry Operators**: "*A is B's lawyer*" → "*B is A's lawyer*"
- 24 types in total

LINGUISTIC OPERATOR	EXAMPLE
NEGATION	NEG(fit into) = not fit into
ANTONYM	ANT(fit into) = contain
PARAPHRASE	PARA(fit into) = put into
PARAPHRASE INVERSION	PARA(ANT(fit into)) = Para(contain) = hold inside
NEGATION ANTONYM	NEG(ANT(fit into)) = NEG(contain) = not contain
NEGATION PARAPHRASE	NEG(PARA(fit into)) = NEG(put into) = not put into
NEGATION PARA_INV	NEG(PARA(ANT(fit into))) = NEG(PARA(contain))= NEG(hold inside) =not hold inside

Experiments

Masked Word Prediction (MWP)

1. BERT / RoBERTa
2. ERNIE (KG-enhanced LM)
3. BART (Seq2seq)

Novel Entity Pair: prindag and fluberg

Masked Word Prediction:
A prindag is lighter than a fluberg, so a prindag should float *[MASK]* than a fluberg. [more] or [less]

Testing Set: **1.6k human-curated**

Evaluation Settings:

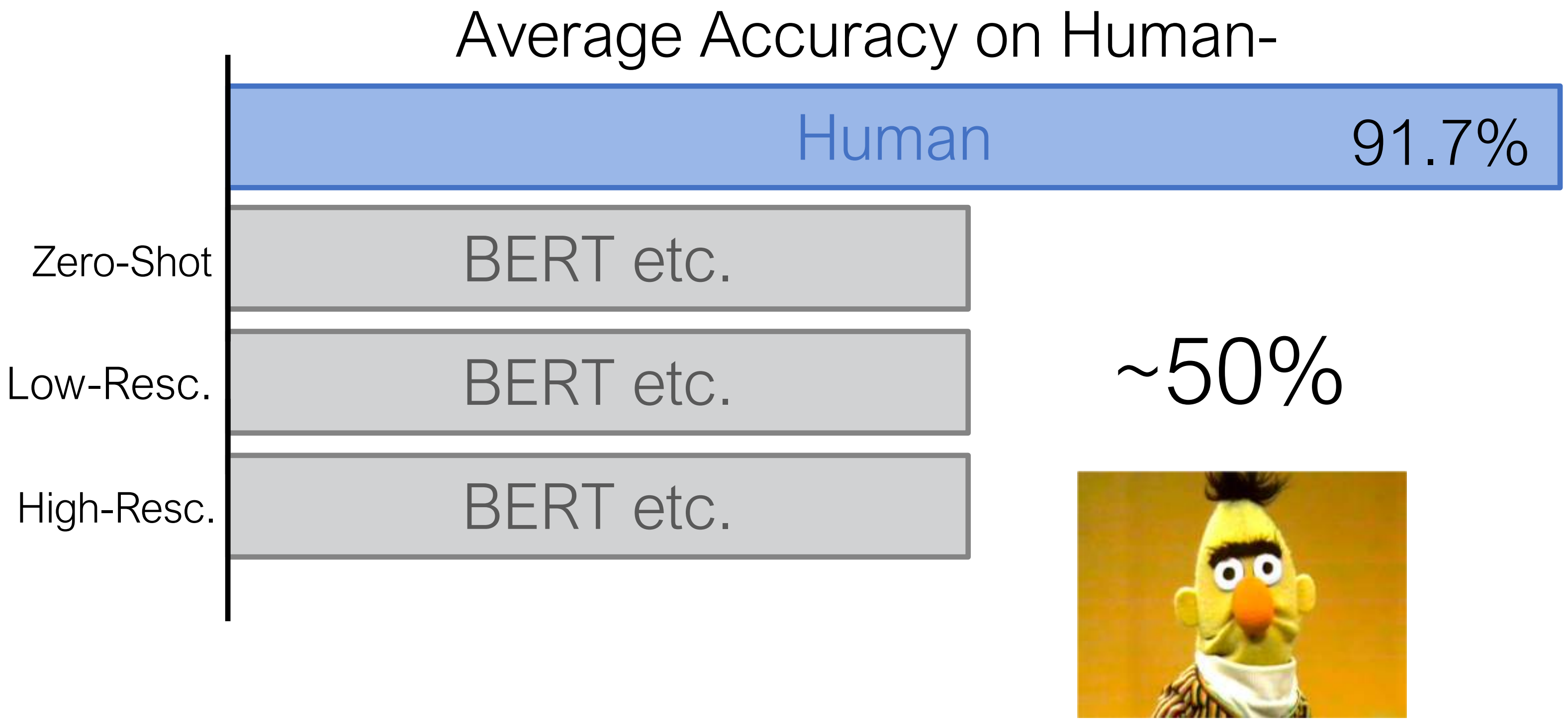
1. **Zero-Shot**: without fine-tuning
2. **Low-Resource**: fine-tune on 1k of all verified probes
3. **High-Resource**: fine-tune on all verified probes (9k)
4. Large-Scale on Raw Data: 100k from the machine generated set

Metric: Average accuracy

Results: Human-Curated Set

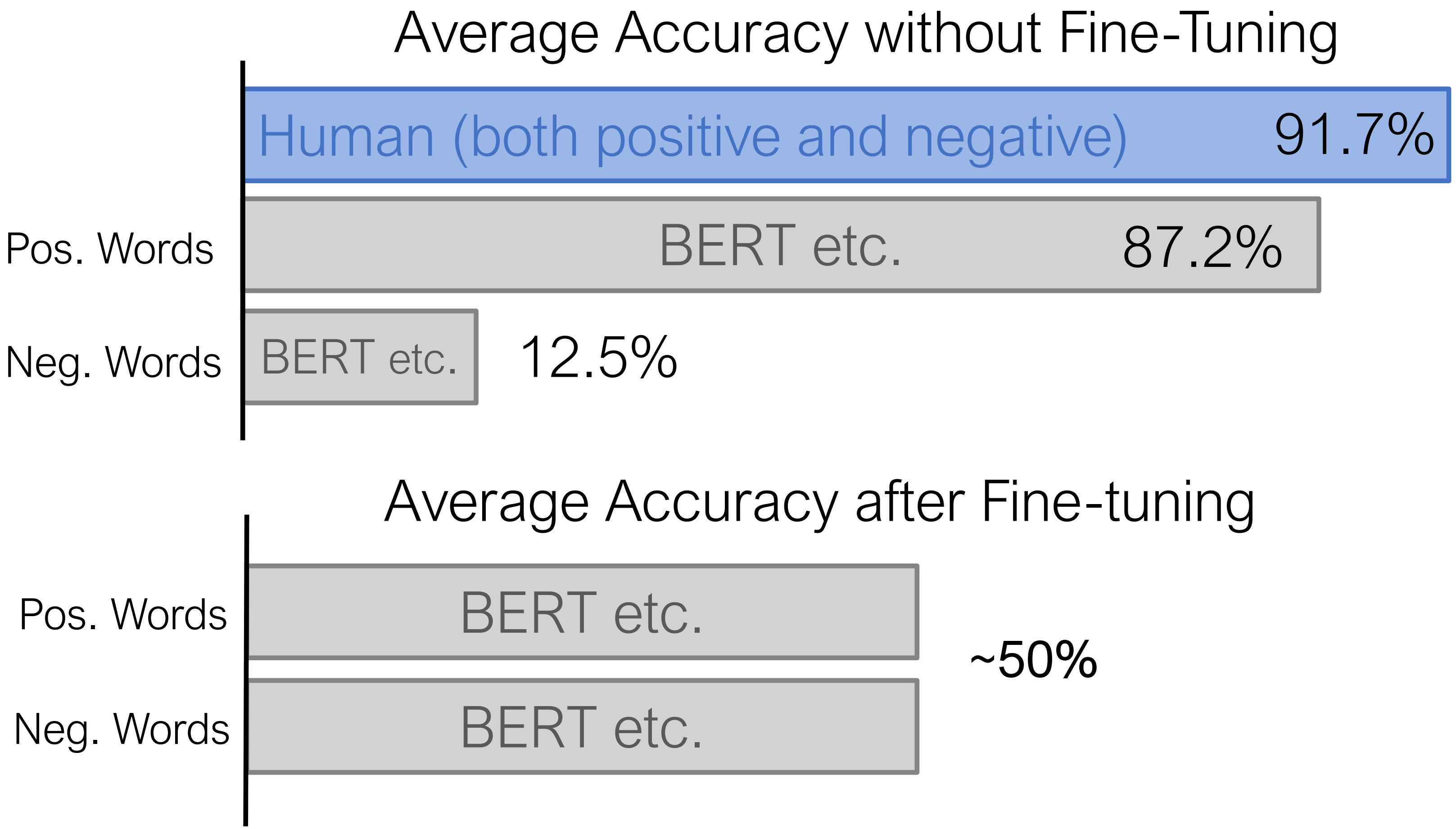
- Random-guessing like performance on all settings for all models.

- Training on similar data does not help achieve real robustness



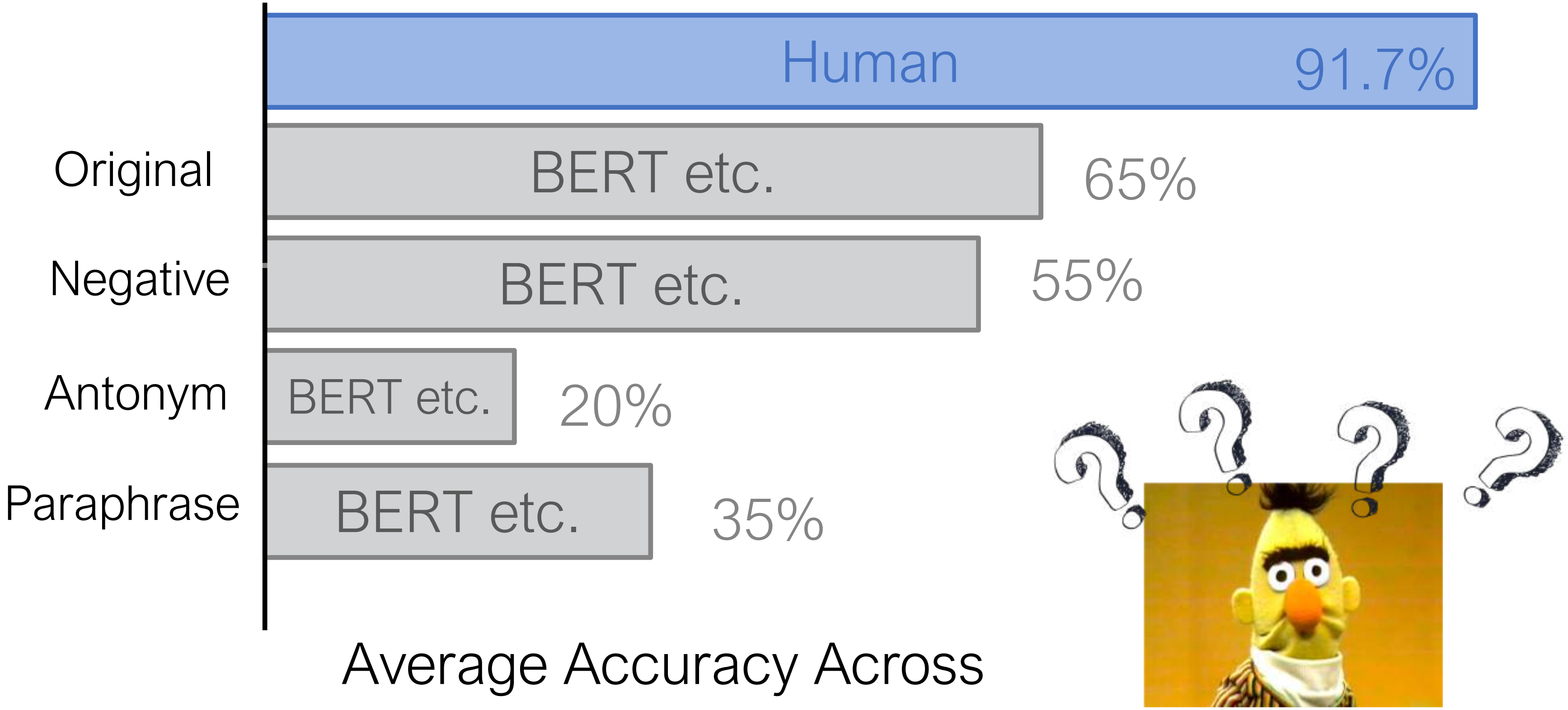
Analysis: Positivity Bias

- Heavy bias towards positive-valence words such as “more”, “better”, “easier”.
- Fine-tuning on RICA mitigates the imbalance issue (but still fails)



Analysis: Robustness Issue

- Severe variation among different linguistic perturbation operators

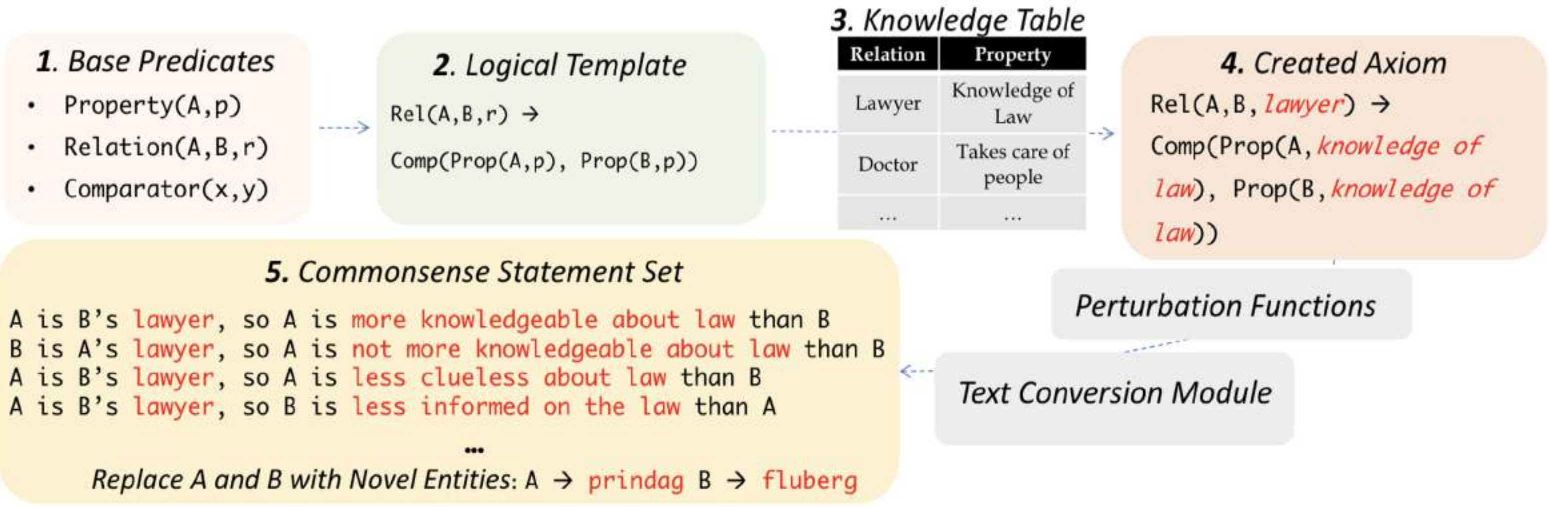


Average Accuracy Across Perturbation Types

Summary for RICA

Combining common sense with information expressed in NL to make inferences

Producing consistent inferences amidst logically-equivalent yet linguistically-varied paraphrases.



https://sites.google.com/usc.edu/rica

Cross-task generalization in NLP

Learning at the **instance-level**

Generalize from a few ***seen training instances***,
to multiple ***unseen test instances***.

Train	This movie is extraordinary.	**Positive**
	Watching it is a waste of time.	**Negative**
Test	It’s such a wonderful movie!	**?**
	I’m so disappointed!	**?**

Task: Movie Review Sentiment Classification

Cross-task generalization in NLP

Learning at the **instance-level**

Generalize from a few ***seen training instances***,
to multiple ***unseen test instances***.

Train	This movie is extraordinary.	**Positive**
	Watching it is a waste of time.	**Negative**
Test	It's such a wonderful movie!	**?**
	I'm so disappointed!	**?**

Task: Movie Review Sentiment Classification

Learning at the **task-level**

Generalize from a few ***seen training tasks***,
to multiple ***unseen test tasks***.

Train	Movie Review Sentiment Classification	Reading Comprehension on News	Biomedical Relation Extraction	...
Test	Paraphrase Identification	Commonsense Multiple-choice QA	Mining Alpha Factors from News Corpora	...

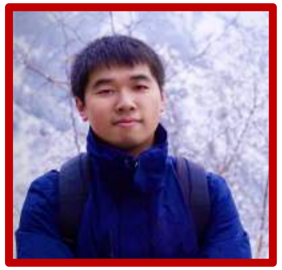
Goal: Achieve competitive performance on the test task with fewer annotations.

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

- 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.

(Ye et al., EMNLP 2021)


Qinyuan Ye


Bill Yuchen Lin


Xiang Ren

CrossFit: Quick Summary

NLP Few-shot Gym 💦

- Gather **160 diverse few-shot tasks** in text-to-text format

Datasets

(Ye et al., EMNLP 2021)

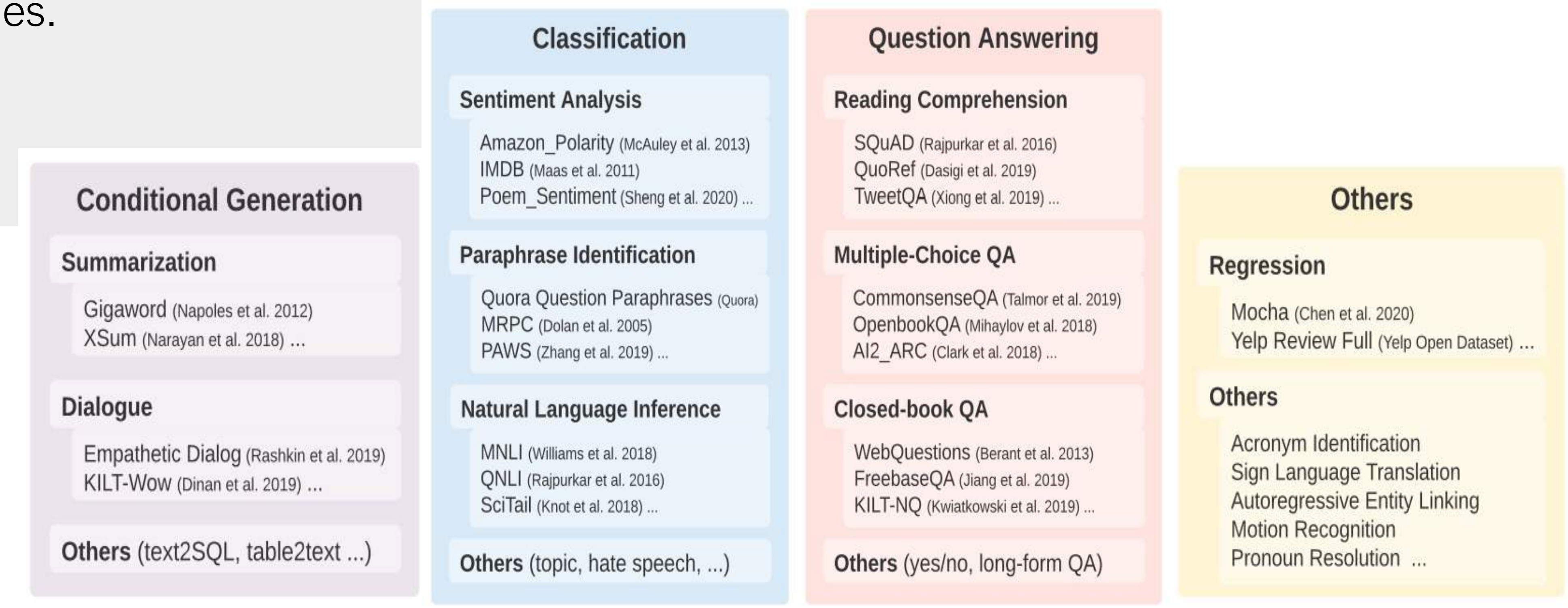
Task Name	Ontology	Reference
acronym_identification	other	Pouran Ben Veyseh et al. 2020
ade_corpus_v2-classification	cls/other	Gurulingappa et al. 2012
ade_corpus_v2-dosage	other/slot filling	Gurulingappa et al. 2012
ade_corpus_v2-effect	other/slot filling	Gurulingappa et al. 2012
adversarialqa	qa/machine reading comprehension	Bartolo et al. 2020
aeslc	cg/summarization	Zhang and Tetreault 2019
ag_news	cls/topic	Gulli (link)
ai2_arc	qa/multiple-choice qa	Clark et al. 2018
amazon_polarity	cls/sentiment analysis	McAuley and Leskovec 2013
anli	cls/nli	Nie et al. 2020
app_reviews	other/regression	Missing
aqua_rat	qa/multiple-choice qa	Ling et al. 2017
art (abductive nli)	other	Bhagavatula et al. 2020
aslg_pc12	other	Othman and Jemni 2012
biomrc	qa/machine reading comprehension	Pappas et al. 2020
blimp-anaphor_gender_agreement	other/linguistic phenomenon	Warstadt et al. 2020
blimp-anaphor_number_agreement	other/linguistic phenomenon	Warstadt et al. 2020
blimp-determiner_noun_agreement_with_adj_irregular_1	other/linguistic phenomenon	Warstadt et al. 2020
blimp-ellipsis_n_bar_1	other/linguistic phenomenon	Warstadt et al. 2020
blimp-ellipsis_n_bar_2	other/linguistic phenomenon	Warstadt et al. 2020
blimp-existential_there_quantifiers_1	other/linguistic phenomenon	Warstadt et al. 2020
blimp-irregular_past_participle_adjectives	other/linguistic phenomenon	Warstadt et al. 2020
blimp-sentential_negation_npi_licensor_present	other/linguistic phenomenon	Warstadt et al. 2020
blimp-sentential_negation_npi_scope	other/linguistic phenomenon	Warstadt et al. 2020
blimp-wh_questions_object_gap	other/linguistic phenomenon	Warstadt et al. 2020
boolq	qa/binary	Clark et al. 2019
break-QDMR	other	Wolfson et al. 2020
break-QDMR-high-level	other	Wolfson et al. 2020
circa	cls/other	Louis et al. 2020
climate_fever	cls/fact checking	Diggelmann et al. 2020
codah	qa/multiple-choice qa	Chen et al. 2019
common_gen	other	Lin et al. 2020b
commonsense_qa	qa/multiple-choice qa	Talmor et al. 2019
cos_e	other/generate explanation	Rajani et al. 2019
cosmos_qa	qa/multiple-choice qa	Huang et al. 2019
crawl_domain	other	Zhang et al. 2020
crows_pairs	other	Nangia et al. 2020
dbpedia_14	cls/topic	Lehmann et al. 2015
definite_pronoun_resolution	other	Rahman and Ng 2012
discovery	cls/other	Sileo et al. 2019
dream	qa/multiple-choice qa	Sun et al. 2019
duorc	qa/machine reading comprehension	Saha et al. 2018
e2e_nlg_cleaned	other	Dušek et al. 2020, 2019
eli5-askh	qa/long-form qa	Fan et al. 2019
eli5-asks	qa/long-form qa	Fan et al. 2019
eli5-eli5	qa/long-form qa	Fan et al. 2019
emo	cls/emotion	Chatterjee et al. 2019
emotion	cls/emotion	Saravia et al. 2018
empathetic_dialogues	cg/dialogue	Rashkin et al. 2019
ethos-directed_vs_generalized	cls/hate speech detection	Mollas et al. 2020
ethos-disability	cls/hate speech detection	Mollas et al. 2020
ethos-gender	cls/hate speech detection	Mollas et al. 2020
ethos-national_origin	cls/hate speech detection	Mollas et al. 2020
ethos-race	cls/hate speech detection	Mollas et al. 2020
ethos-religion	cls/hate speech detection	Mollas et al. 2020
ethos-sexual_orientation	cls/hate speech detection	Mollas et al. 2020
financial_phrasebank	cls/sentiment analysis	Malo et al. 2014
freebase_qa	qa/closed-book qa	Jiang et al. 2019
gigaword	cg/summarization	Napoles et al. 2012
glue-cola	cls/other	Warstadt et al. 2019
glue-mnli	cls/nli	Williams et al. 2018
glue-mrpc	cls/paraphrase	Dolan and Brockett 2005
glue-qnli	cls/nli	Rajpurkar et al. 2016
glue-qqp	cls/paraphrase	(link)
glue-rte	cls/nli	Dagan et al. 2005; Bar-Haim et al. 2006 Giampiccolo et al. 2007; Bentivogli et al. 2009
glue-sst2	cls/sentiment analysis	Socher et al. 2013
glue-wnli	cls/nli	Levesque et al. 2012
google_wellformed_query	cls/other	Faruqui and Das 2018
hate_speech18	cls/hate speech detection	de Gibert et al. 2018
hate_speech_offensive	cls/hate speech detection	Davidson et al. 2017
hatexplain	cls/hate speech detection	Mathew et al. 2020
health_fact	cls/fact checking	Kotonya and Toni 2020
hellaswag	qa/multiple-choice qa	Zellers et al. 2019
hotpot_qa	qa/machine reading comprehension	Yang et al. 2018
imdb	cls/sentiment analysis	Maas et al. 2011
jeopardy	qa/closed-book qa	(link)
kilt_ay2	other/entity linking	Hoffart et al. 2011

Continued on next page

CrossFit: Quick Summary

NLP Few-shot Gym 💦

- Gather **160 diverse few-shot tasks** in text-to-text format
- Manually **group the tasks** into categories and sub-categories.

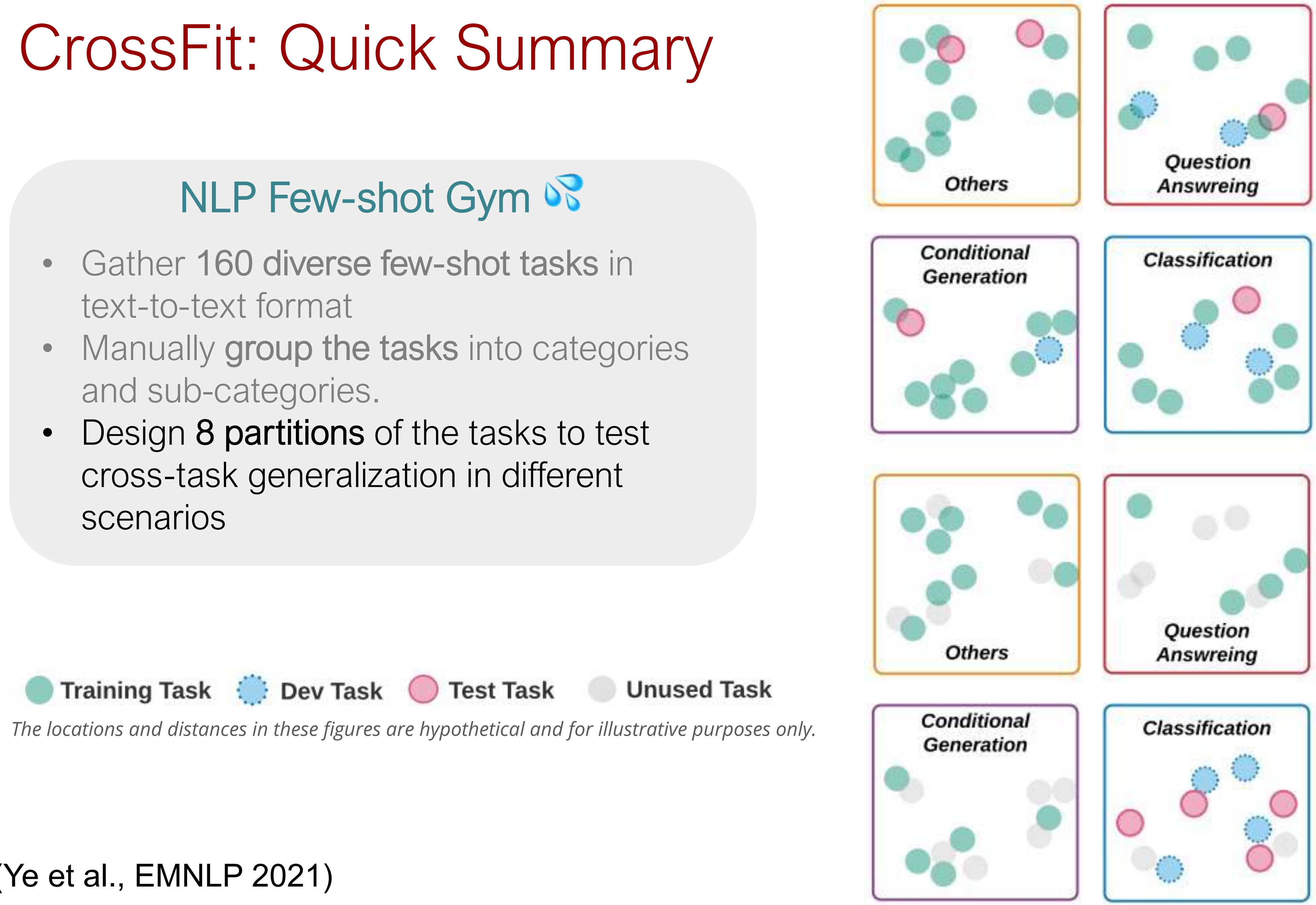


(Ye et al., EMNLP 2021)

CrossFit: Quick Summary

NLP Few-shot Gym 💦

- Gather **160 diverse few-shot tasks** in text-to-text format
- Manually **group the tasks** into categories and sub-categories.
- Design **8 partitions** of the tasks to test cross-task generalization in different scenarios



The locations and distances in these figures are hypothetical and for illustrative purposes only.

Partition 1: Random
Randomly split 160 tasks into 120/20/20 for train/dev/test tasks.

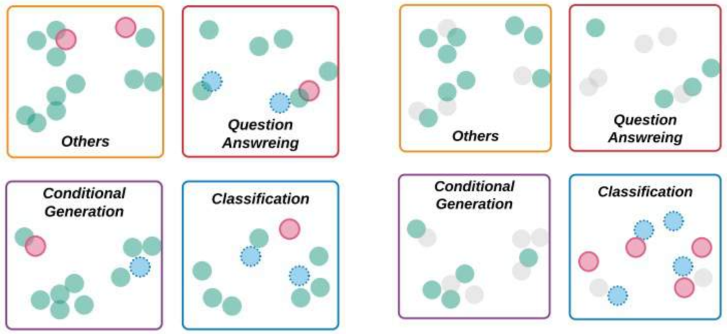
Partition 2.1: 45non-class
Train: 45 non-classification tasks
Dev/Test: 10 classification tasks

(Ye et al., EMNLP 2021)

CrossFit: Quick Summary

NLP Few-shot Gym

- Gather **160 diverse few-shot tasks** in text-to-text format
- Manually **group the tasks** into categories and sub-categories.
- Design **8 partitions** of the tasks to test cross-task generalization in different scenarios



(Ye et al., EMNLP 2021)

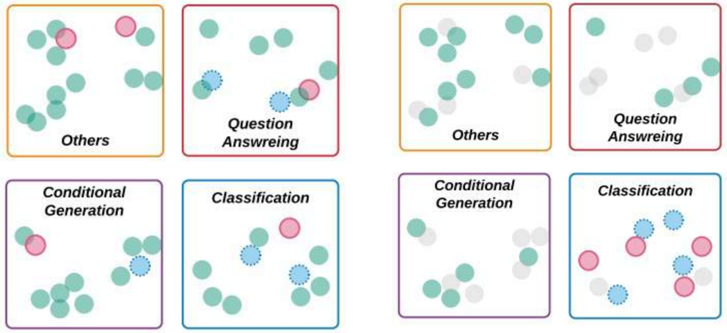
CrossFit Setting

Large-scale Pre-training

CrossFit: Quick Summary

NLP Few-shot Gym 💦

- Gather **160 diverse few-shot tasks** in text-to-text format
- Manually **group the tasks** into categories and sub-categories.
- Design **8 partitions** of the tasks to test cross-task generalization in different scenarios



CrossFit 🏋️ Setting

Large-scale Pre-training

\+ Upstream Learning on a set of seen tasks (T_{train})

Using multi-task learning and meta-learning methods (e.g., MAML, Reptile)

(Ye et al., EMNLP 2021)

CrossFit: Quick Summary

NLP Few-shot Gym

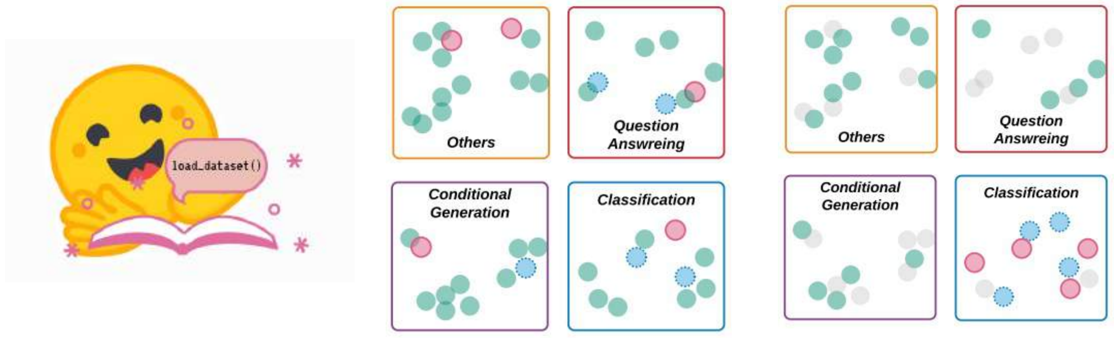
- Gather **160 diverse few-shot tasks** in text-to-text format
- Manually **group the tasks** into categories and sub-categories.
- Design **8 partitions** of the tasks to test cross-task generalization in different scenarios

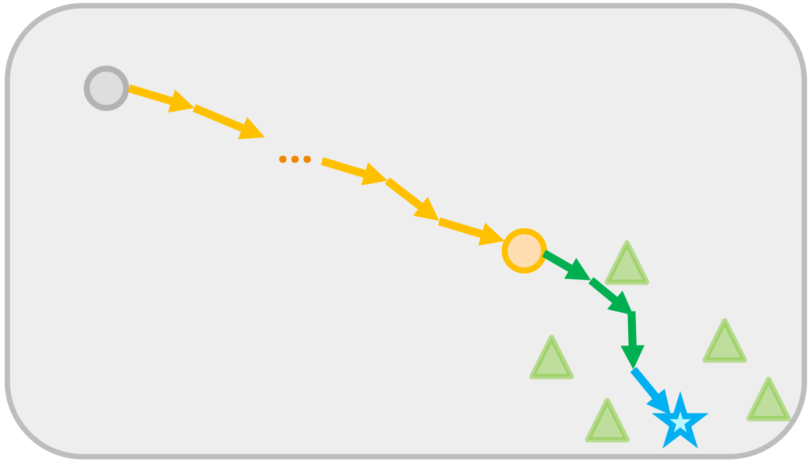
CrossFit Setting

Large-scale Pre-training

+ Upstream Learning on a set of seen tasks (T_{train})

+ Downstream Fine-tuning on an unseen target task (T_{test})





Model Parameter Space

(Ye et al., EMNLP 2021)

Evaluation Metric

- We define ***Average Relative Gain*** (ARG), to measure the overall performance gain on all unseen tasks.
- ARG is the relative performance changes before and after the upstream learning stage for each test task, and averaged across all test tasks.
- ***This is not a perfect metric***, but it helps us to get a general sense. We still plot and report relative gain for individual tasks.

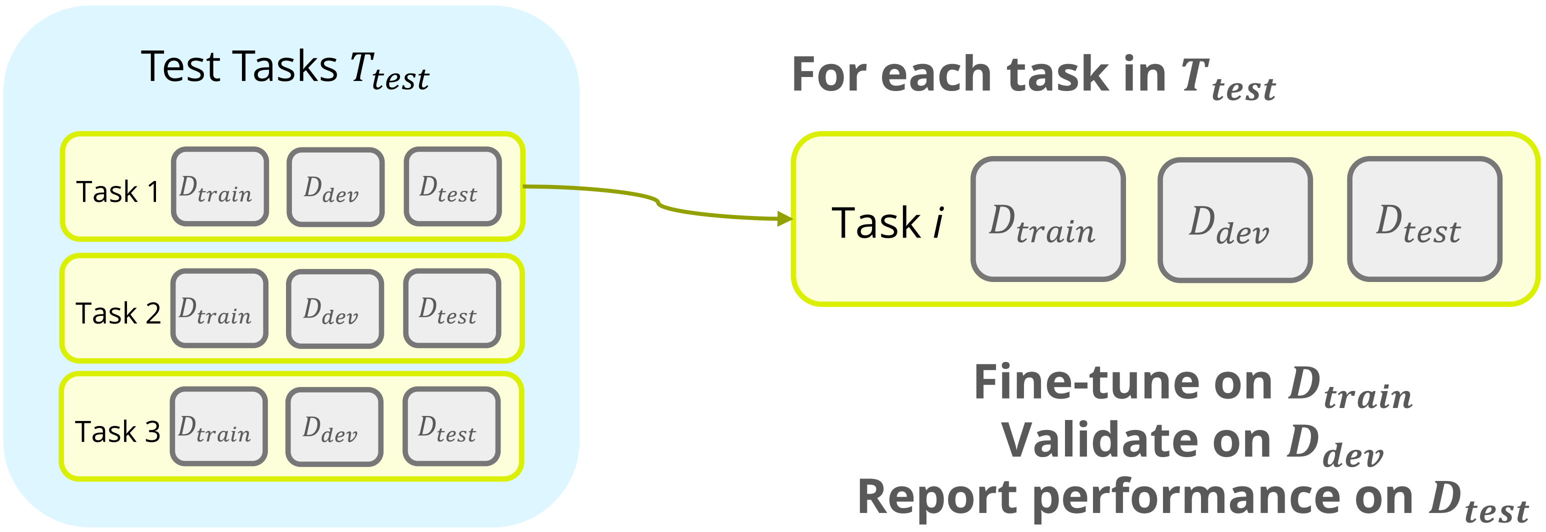
Example

(40%-25%) /2=7.5%

	Direct FT	***Upstream + FT***	***Rel. Gain***	***ARG***
Task A	50% F1	70% F1	40%	7.5%
Task B	40% Acc.	30% Acc.	-25%	

Experiments

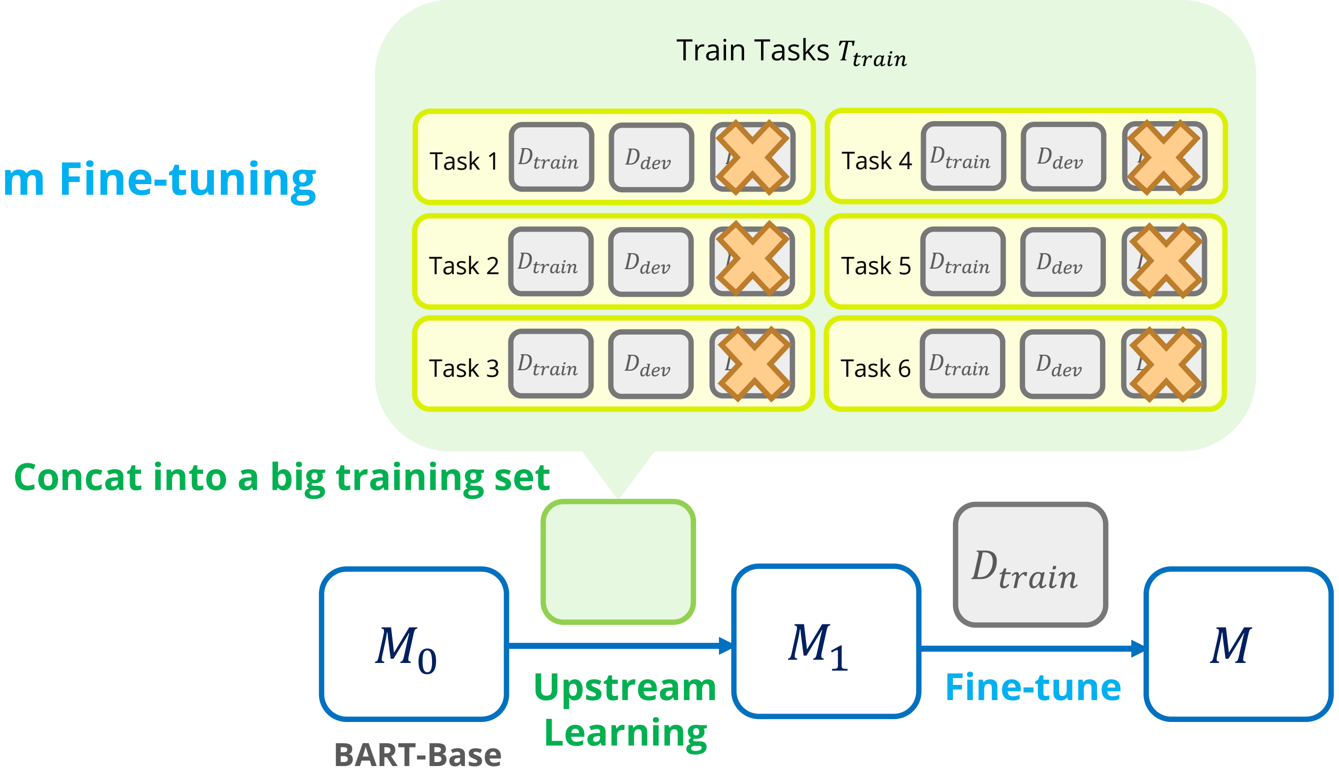
- We mainly use **BART-Base** (Lewis et al., 2020) as the main model for our analysis.
  - Also we verify some of our findings with **BART-Large** and **T5-v1.1-Base** (Raffel et al., 2019)
- Methods for comparison
  - **Downstream Fine-tuning** (also used as the baseline for computing ARG)



Experiments

- We mainly use **BART-Base** (Lewis et al., 2020) as the main model for our analysis.
  - Also we verify some of our findings with **BART-Large** and **T5-v1.1-Base** (Raffel et al., 2019)

- Methods for comparison
  - **Downstream Fine-tuning**
  - **Upstream Learning then Downstream Fine-tuning**
    - Multi-task Learning

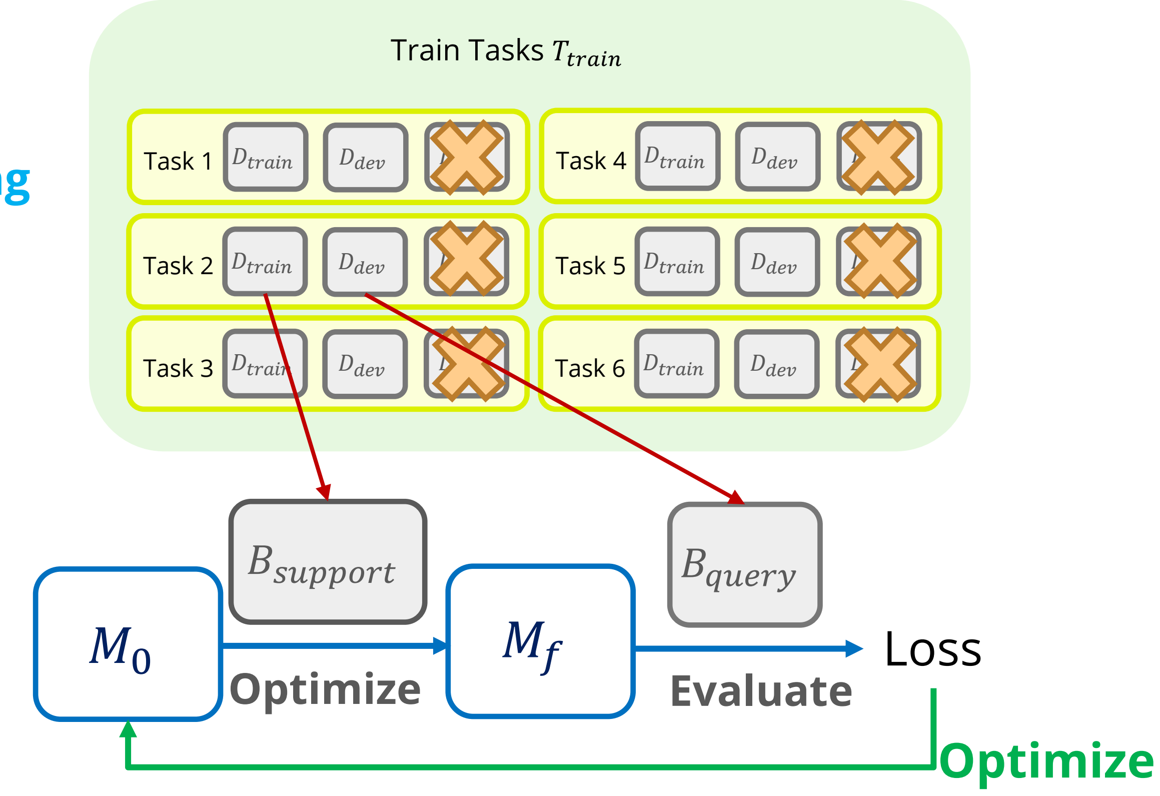


Experiments

- We mainly use **BART-Base** (Lewis et al., 2020) as the main model for our analysis.
    - Also we verify some of our findings with **BART-Large** and **T5-v1.1-Base** (Raffel et al., 2019)
- Methods for comparison
    - **Downstream Fine-tuning**
    - **Upstream Learning then Downstream Fine-tuning**
        - Multi-task Learning
        - Model Agnostic Meta-learning (Finn et al., 2017)
        - First-order MAML
        - Reptile (Nichol et al., 2017)

Variants of MAML

One update in upstream learning with MAML



Question 1
Is upstream learning using seen tasks helpful?

Method
We applied multi-task learning and meta-learning algorithms during upstream learning.

Findings

Yes! Upstream learning methods do help LMs to acquire cross-task generalization.

The conclusion holds on different splits of seen/unseen tasks, and with different upstream learning methods.

Evidence 1

ARG (defined earlier) is ***positive*** for all 8 partitions and all 4 upstream learning methods

No.	Shorthand	ARG(Multi)	ARG(MAML)	ARG(FoMAML)	ARG(Rept.)
1	Random	35.06%	28.50%	22.69%	25.90%
2.1	45cls	11.68%	9.37%	10.28%	13.36%
2.2	23cls+22non-cls	11.82%	9.69%	13.75%	14.34%
2.3	45non-cls	11.91%	9.33%	11.20%	14.14%
3.1	Held-out-NLI	16.94%	12.30%	12.33%	14.46%
3.2	Held-out-Para	18.21%	17.90%	21.57%	19.72%
4.1	Held-out-MRC	32.81%	27.28%	28.85%	28.85%
4.2	Held-out-MCQA	12.20%	4.69%	6.73%	7.67%

Evidence 2

When we aggregate test performance gain from all upstream learning methods and partitions...

⬆	***>5% relative gain***	***51.47%***
➡	***within ±5%***	***35.93%***
⬇	***<-5% relative gain***	***12.60%***

Question 2

How does the selection of seen tasks influence the performance?

Method – Controlled Experiments

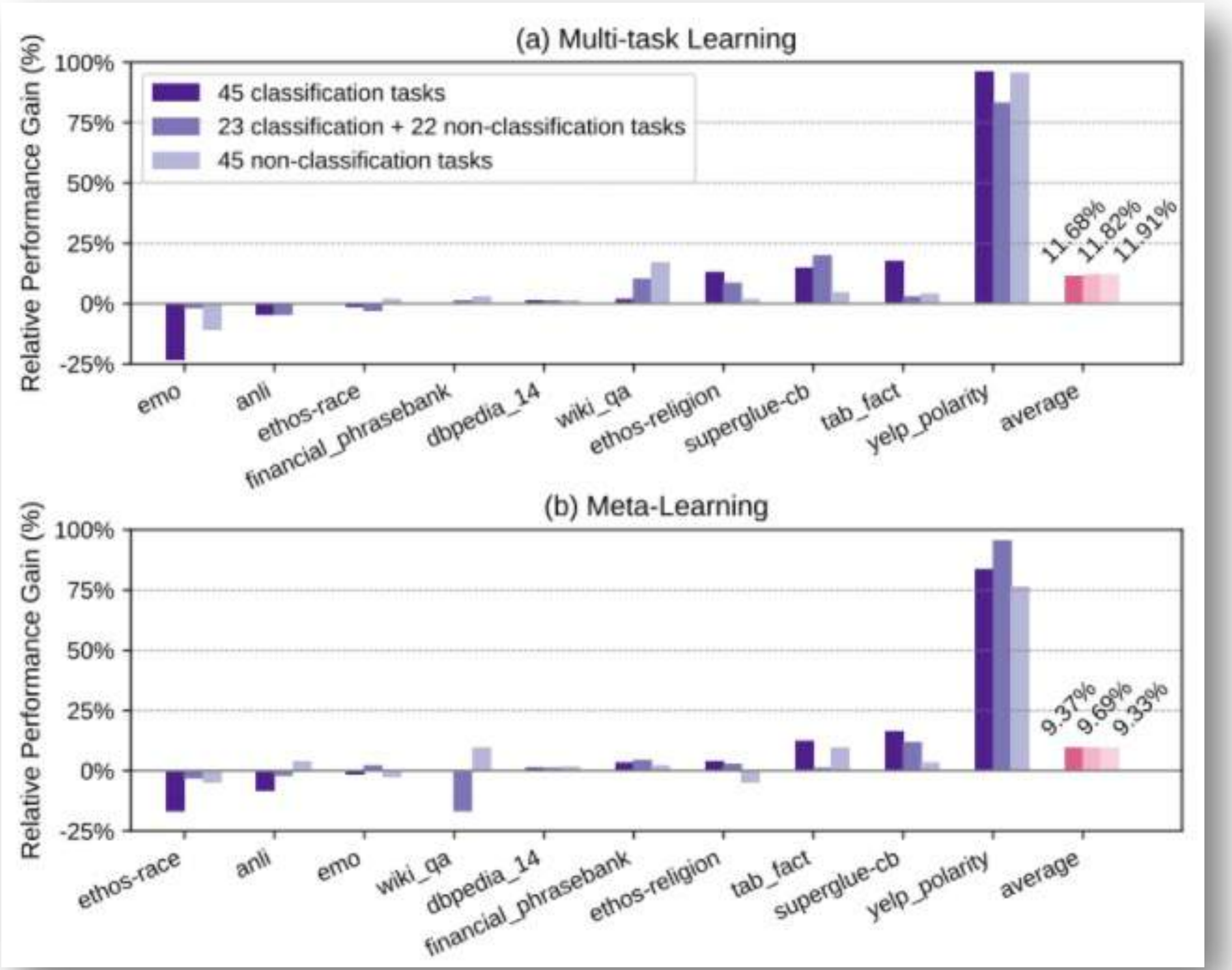
Seen tasks: (1) 100% classification
(2) 50% class + 50% non-class
(3) 100% non-classification

Unseen tasks: 100% classification

Findings

Classification tasks and non-classification tasks seem to be equivalently helpful.

Our understanding of tasks may not align with how models learn transferable skills!



Bar height: relative performance gain (ARG) with vs. without upstream learning

Question 3

Does the improved cross-task generalization ability go beyond few-shot settings?

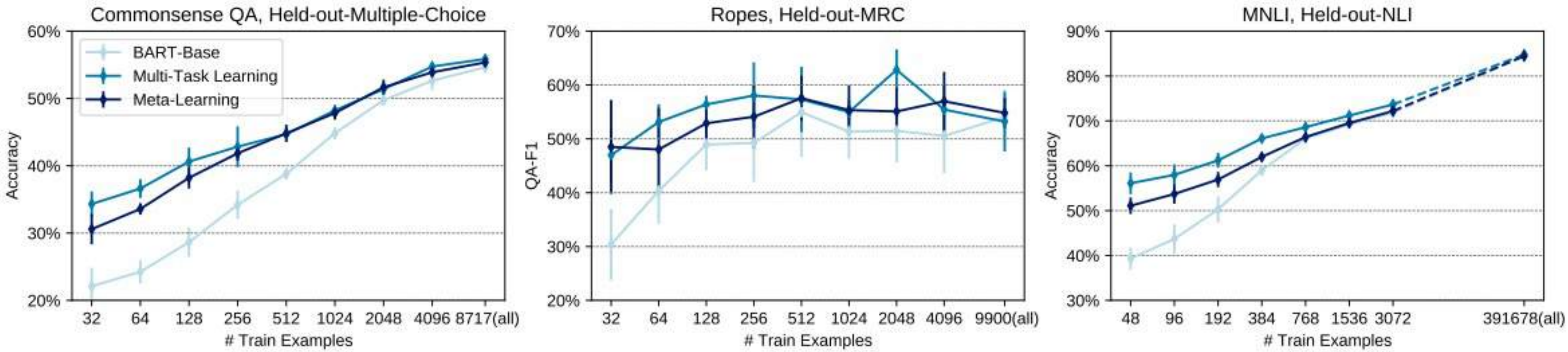
Method

Increase the amount of training data for downstream/unseen tasks (32, 64, → 4.1k, 8.7k)

Findings

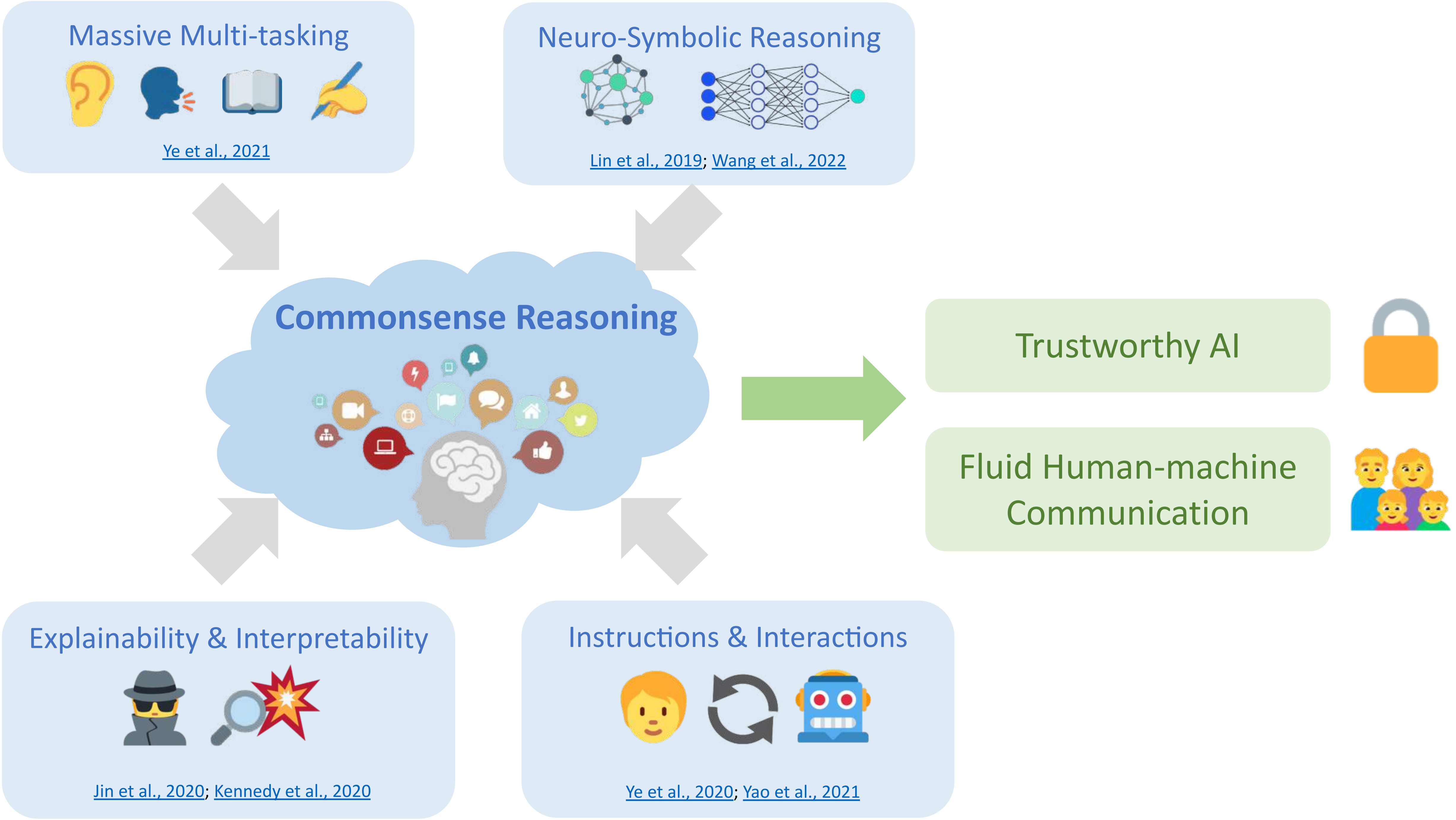
Cross-task generalization helps most on CommonsenseQA, ROPES and MNLI.

On these three datasets, the benefits brought by upstream learning methods extend into medium resource cases with up to 2048 training examples.



- We found that …
  - **Upstream learning methods** such as multi-task learning and meta-learning help pre-trained LMs to **acquired cross-task generalization**.
  - Task similarity in terms of task format ***does not*** align with how models learn transferable skills.

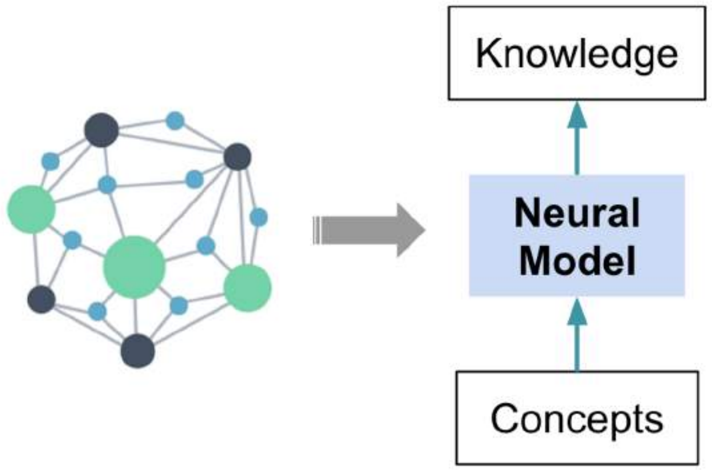
- We envision the **CrossFit** 🏋️ Challenge and the **NLP Few-shot Gym** 💦 to serve as the testbed for many interesting "meta-problems"
  - Generating Prompts? (Shin et al., 2020; Gao et al., 2020)
  - Select appropriate upstream tasks? (Zamir et al., 2018; Standley et al., 2020; Vu et al., 2020)
  - Apply task augmentation? (Murty et al., 2021)
  - Continual Learning? (Jin et al., 2021)
  - Task decomposition? (Andreas et al., 2016; Khot et al., 2021)



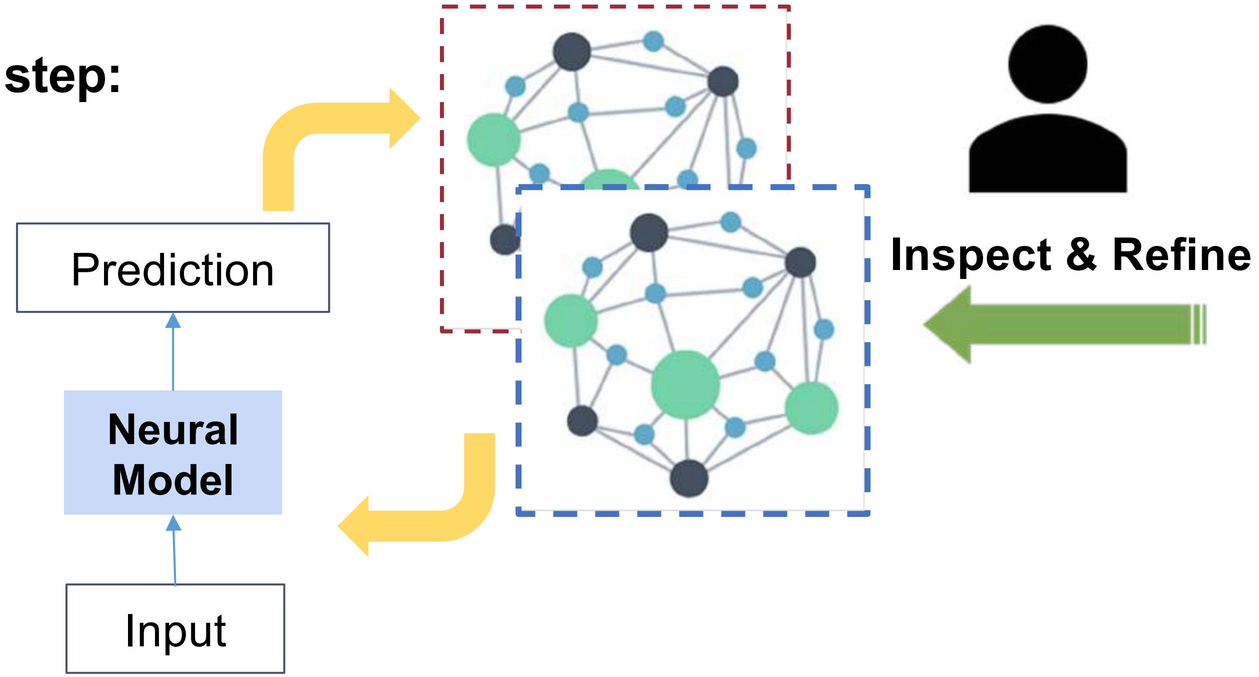
Massive Multi-tasking
Ye et al., 2021
Neuro-Symbolic Reasoning
Lin et al., 2019; Wang et al., 2022
Commonsense Reasoning
Trustworthy AI
Fluid Human-machine Communication
Explainability & Interpretability
Jin et al., 2020; Kennedy et al., 2020
Instructions & Interactions
Ye et al., 2020; Yao et al., 2021

Symbolic knowledge helps create trustworthy NLP models

Prior work:



Next step:



Adding knowledge

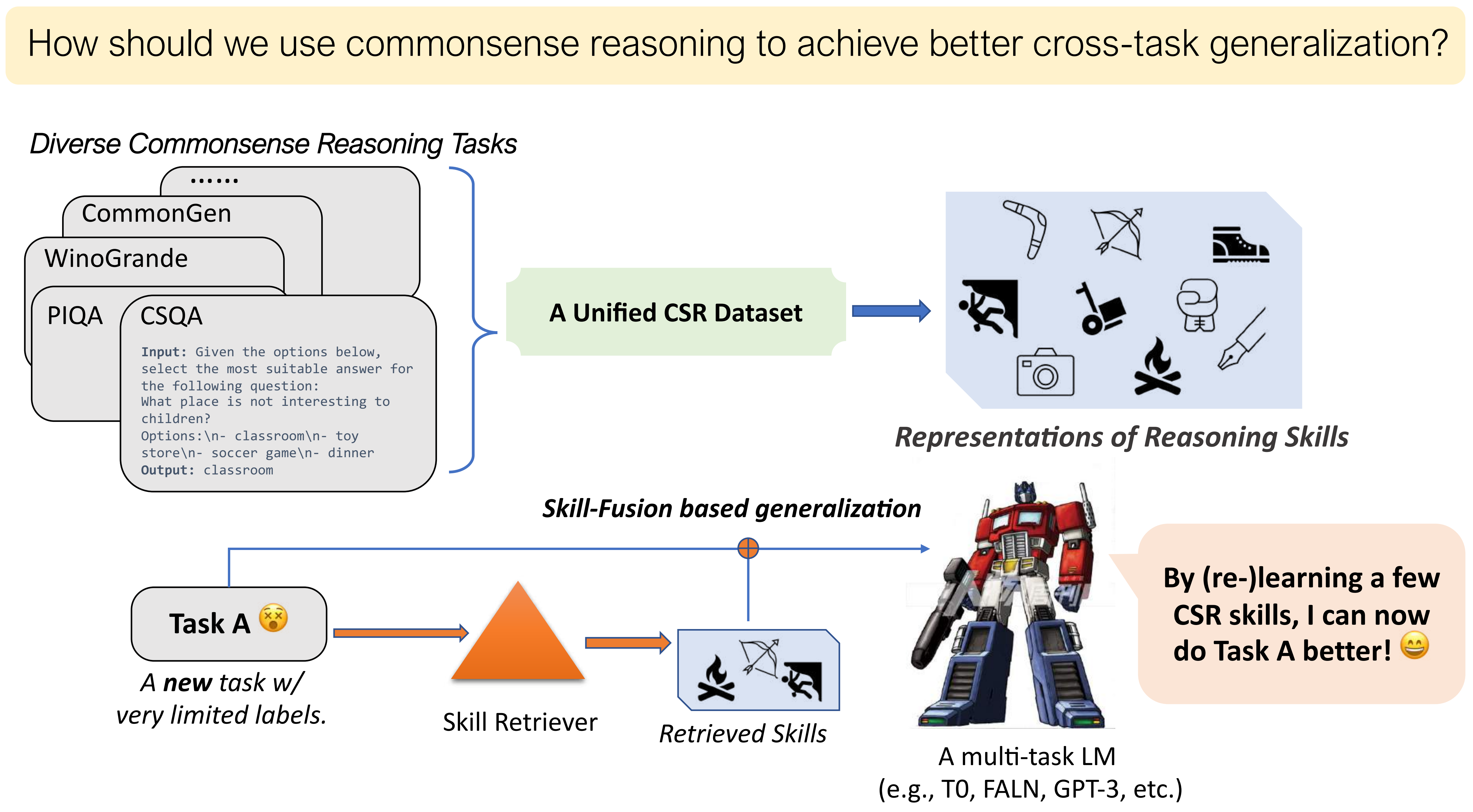
1. + Path for CSQA (EMNLP'20 Findings)
2. + Triplets for KG completion (ACL'21 Findings)
3. + Graph for GCSR (ICLR'22)

Symbolic knowledge as the backbone of model explanation

"Why does a model make a particular decision?"

knowledge for refining model for continual learning

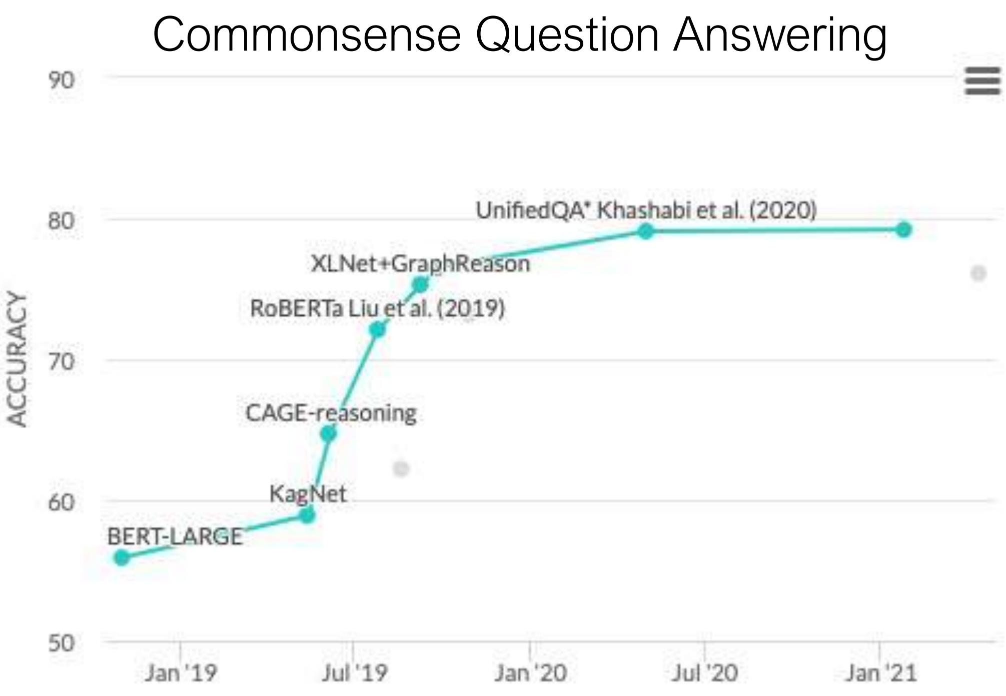
"Can we debug a model?"


How should we use commonsense reasoning to achieve better cross-task generalization?
Diverse Commonsense Reasoning Tasks
......
CommonGen
WinoGrande
PIQA
CSQA
Input: Given the options below, select the most suitable answer for the following question:
What place is not interesting to children?
Options:\n- classroom\n- toy store\n- soccer game\n- dinner
Output: classroom
A Unified CSR Dataset
Representations of Reasoning Skills
Skill-Fusion based generalization
Task A
A new task w/ very limited labels.
Skill Retriever
Retrieved Skills
A multi-task LM
(e.g., T0, FALN, GPT-3, etc.)
By (re-)learning a few CSR skills, I can now do Task A better!

Questions?

Solving a Commonsense Reasoning Dataset

Goal: Perform well on a test set



Paper With Code: CommonsenseQA 1.1

Solving Commonsense Reasoning

Goal: Satisfy the real-world needs

well-rounded

learns fast

robust to variations

People
~~A person~~ performing in front of
people ~~might be nervous~~
find it hard to relax

data efficient

can resolve ambiguity

when is the super bowl | Search

Do you mean *when is the super bowl 2022?*

And more...