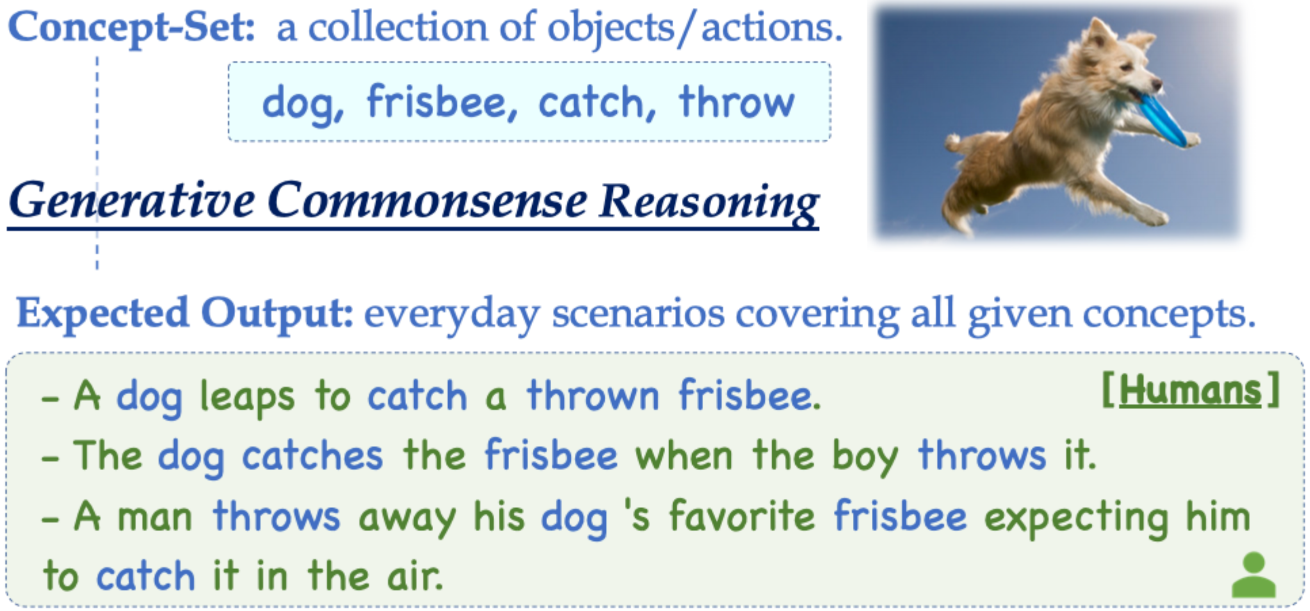
CommonGen: A Constrained Text Generation Challenge for Generative Commonsense Reasoning

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**.



Concept-Set: a collection of objects/actions.

dog, frisbee, catch, throw

Generative Commonsense Reasoning

Expected Output: everyday scenarios covering all given concepts.

- A dog leaps to catch a thrown frisbee.
- The dog catches the frisbee when the boy throws it.
- A man throws away his dog 's favorite frisbee expecting him to catch it in the air.

Input:

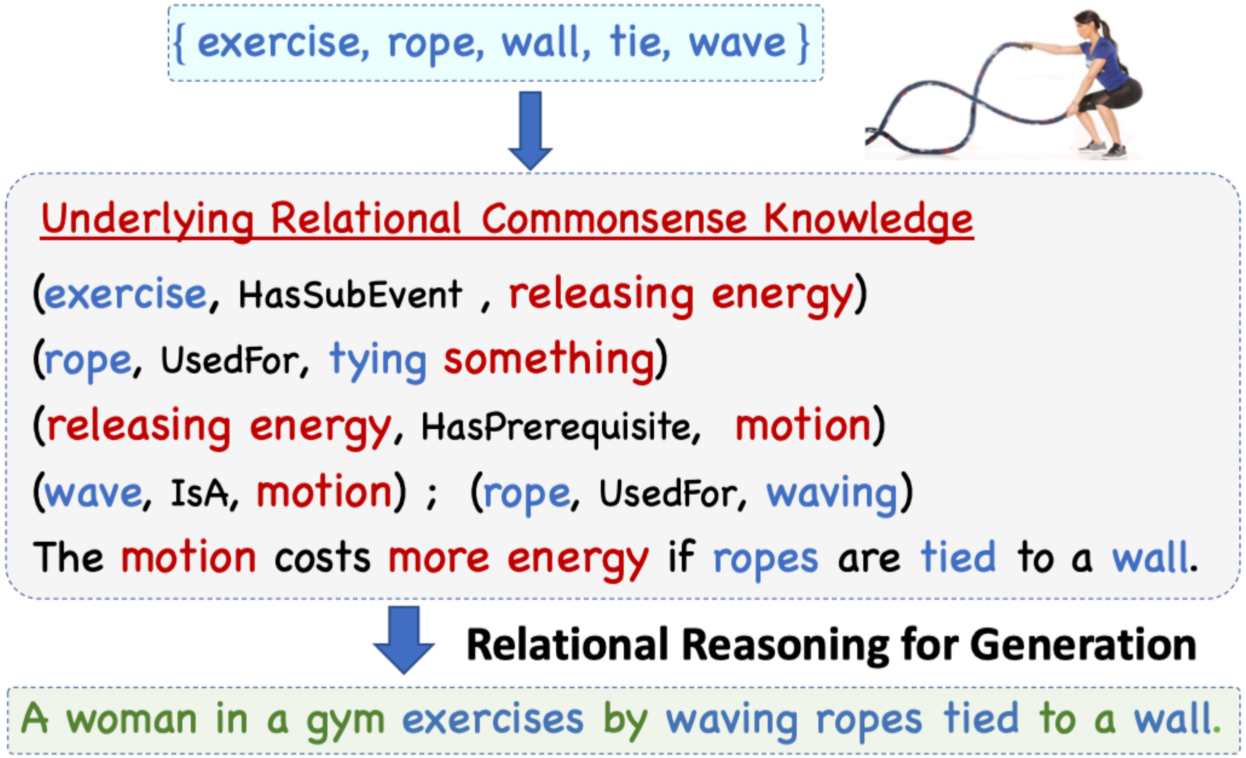
-A set of common concepts (actions & objects)

Output:

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

Why is it hard?
Two key Challenges of CommonGen

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



{ exercise, rope, wall, tie, wave }

Underlying Relational Commonsense Knowledge

(exercise, HasSubEvent , releasing energy)

(rope, UsedFor, tying something)

(releasing energy, HasPrerequisite, motion)

(wave, IsA, motion) ; (rope, UsedFor, waving)

The motion costs more energy if ropes are tied to a wall.

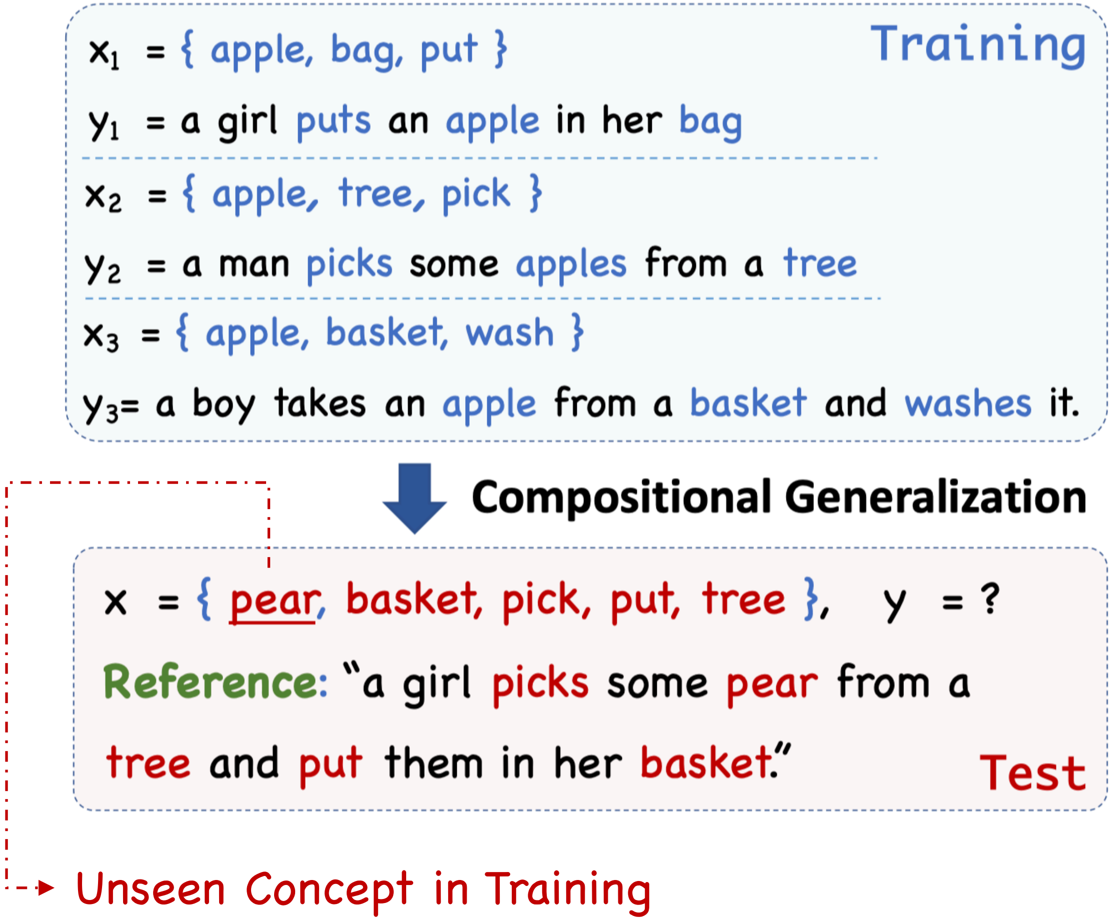
Relational Reasoning for Generation

A woman in a gym exercises by waving ropes tied to a wall.

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 it hard?
Two key Challenges of CommonGen

(2) Compositional Generalization for unseen concept compounds.



Training

x_1 = { apple, bag, put }

y_1 = a girl puts an apple in her bag

x_2 = { apple, tree, pick }

y_2 = a man picks some apples from a tree

x_3 = { apple, basket, wash }

y_3= a boy takes an apple from a basket and washes it.

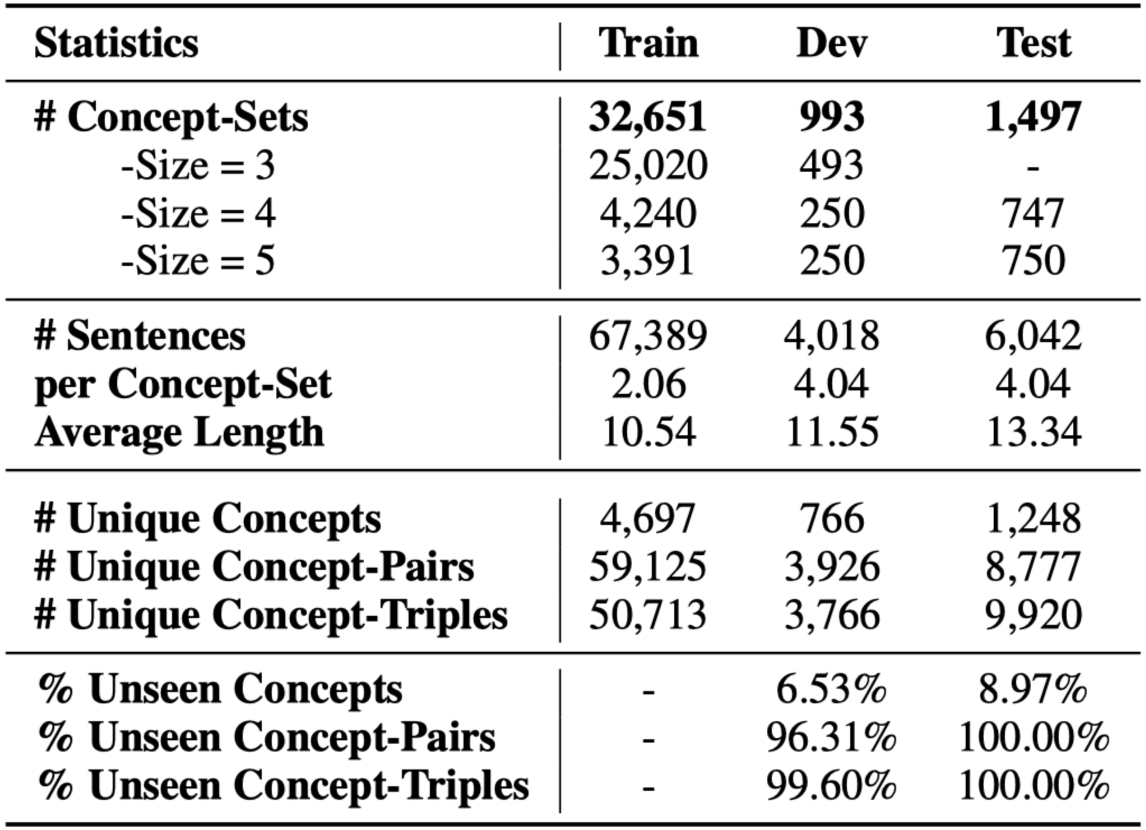
Compositional Generalization

x = { pear, basket, pick, put, tree }, y = ?

Reference: "a girl picks some pear from a tree and put them in her basket."

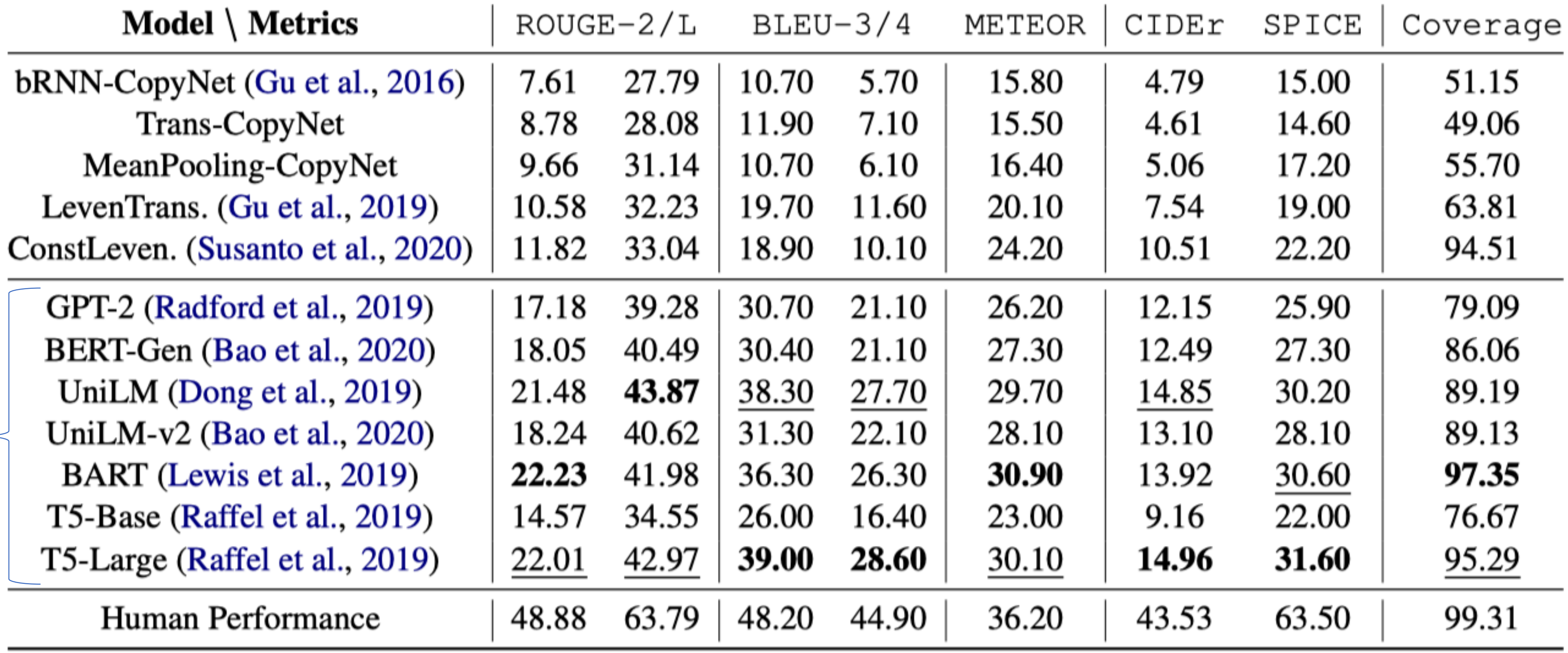
Test

Unseen Concept in Training



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
# Sentences	67,389	4,018	6,042
per Concept-Set	2.06	4.04	4.04
Average Length	10.54	11.55	13.34
# Unique Concepts	4,697	766	1,248
# Unique Concept-Pairs	59,125	3,926	8,777
# Unique Concept-Triples	50,713	3,766	9,920
% Unseen Concepts	-	6.53%	8.97%
% Unseen Concept-Pairs	-	96.31%	100.00%
% Unseen Concept-Triples	-	99.60%	100.00%

Experimental Results



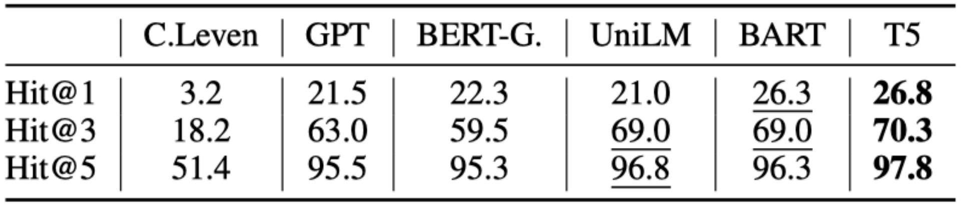
Model \ Metrics	ROUGE-2/L		BLEU-3/4		METEOR	CIDEr	SPICE	Coverage
bRNN-CopyNet (Gu et al., 2016)	7.61	27.79	10.70	5.70	15.80	4.79	15.00	51.15
Trans-CopyNet	8.78	28.08	11.90	7.10	15.50	4.61	14.60	49.06
MeanPooling-CopyNet	9.66	31.14	10.70	6.10	16.40	5.06	17.20	55.70
LevenTrans. (Gu et al., 2019)	10.58	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

Manual Eval.



	C.Leven	GPT	BERT-G.	UniLM	BART	T5
Hit@1	3.2	21.5	22.3	21.0	26.3	**26.8**
Hit@3	18.2	63.0	59.5	69.0	69.0	**70.3**
Hit@5	51.4	95.5	95.3	96.8	96.3	**97.8**

Case Study & Transfer Learning

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.

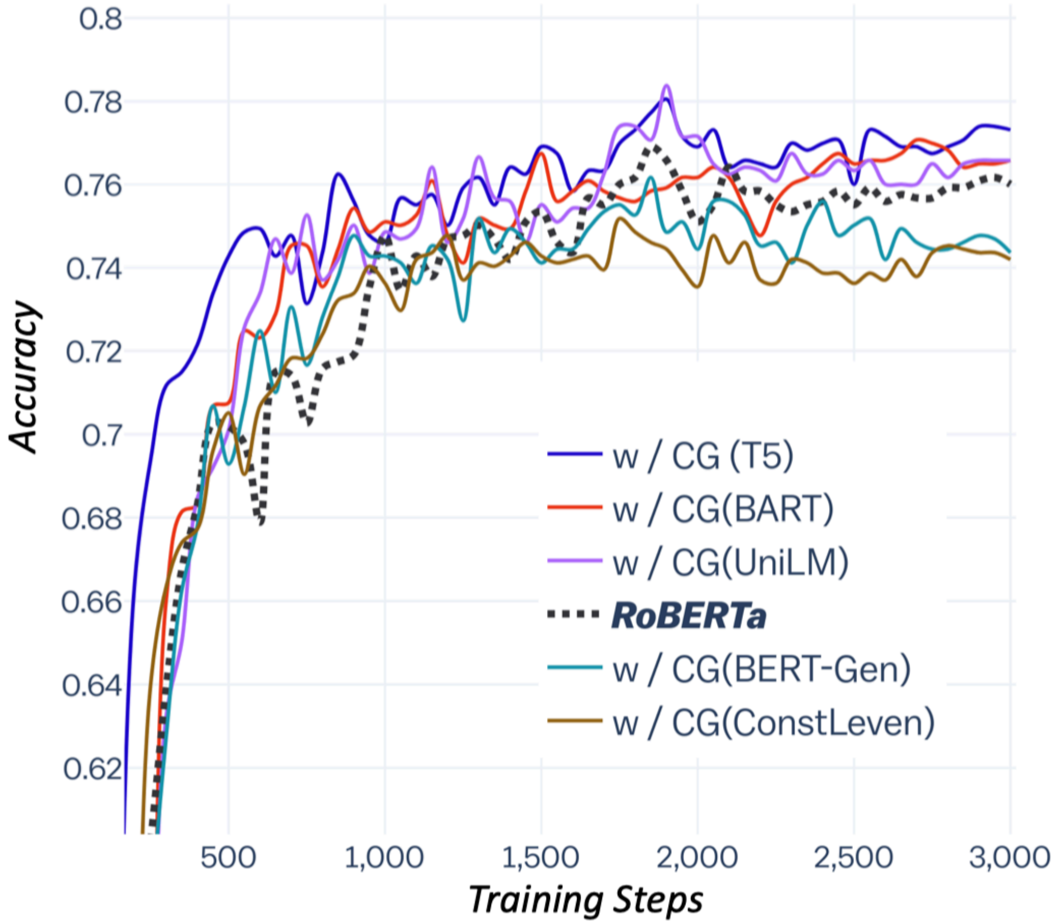
[BERT-Gen]: a woman washes her hands with a sink of soaps.

[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.



Learning curve for the transferring study (acc on dev). We use trained CommonGen models to generate choice-specific context for the CommonsenseQA task.

Thank you for listening!

- Full Paper (non-archival) :
  - https://yuchenlin.xyz/commongen_akbc20.pdf
- Project Page:
  - https://inklab.usc.edu/CommonGen/
- Email yuchen.lin@usc.edu for more questions!