



---

ACL 2022 Tutorial (T3)

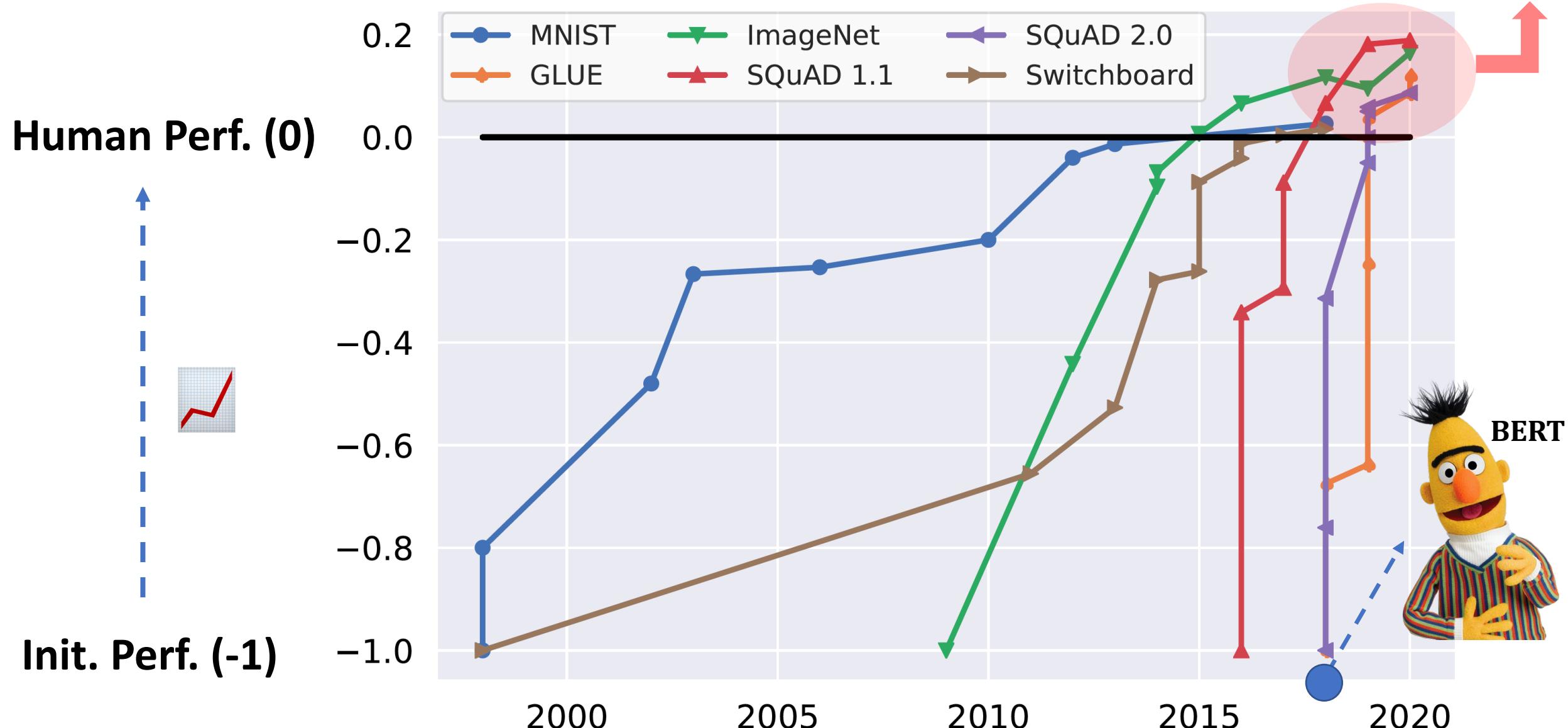
# Incorporating Commonsense Knowledge into Neural NLP Models



Xiang Ren & Bill Yuchen Lin

University of Southern California

[https://inklab.usc.edu/acl22\\_tutorial.\[pdf/pptx\]](https://inklab.usc.edu/acl22_tutorial.[pdf/pptx])



Initial baseline performance and human performance are normalized to **-1** and **0** respectively (Credit: [Kiela et al., 2021](#)).

# Modern NLP models still lack commonsense knowledge.

≡ Google Translate

Text Websites

ENGLISH CHINESE (SIMPLIFIED)

The fruit flies like a banana. ×

果子像香蕉一样飞来飞去。 × ☆

The/ fruit flies/ like/ a banana/. ✓  
The fruit/ flies/ like a banana/. ×

Yuchen PER will travel to NYC LOC ✓ for the ACL ORG conference .

Yuchen PER will travel to ZYC ORG ✗ for the ACL ORG conference .

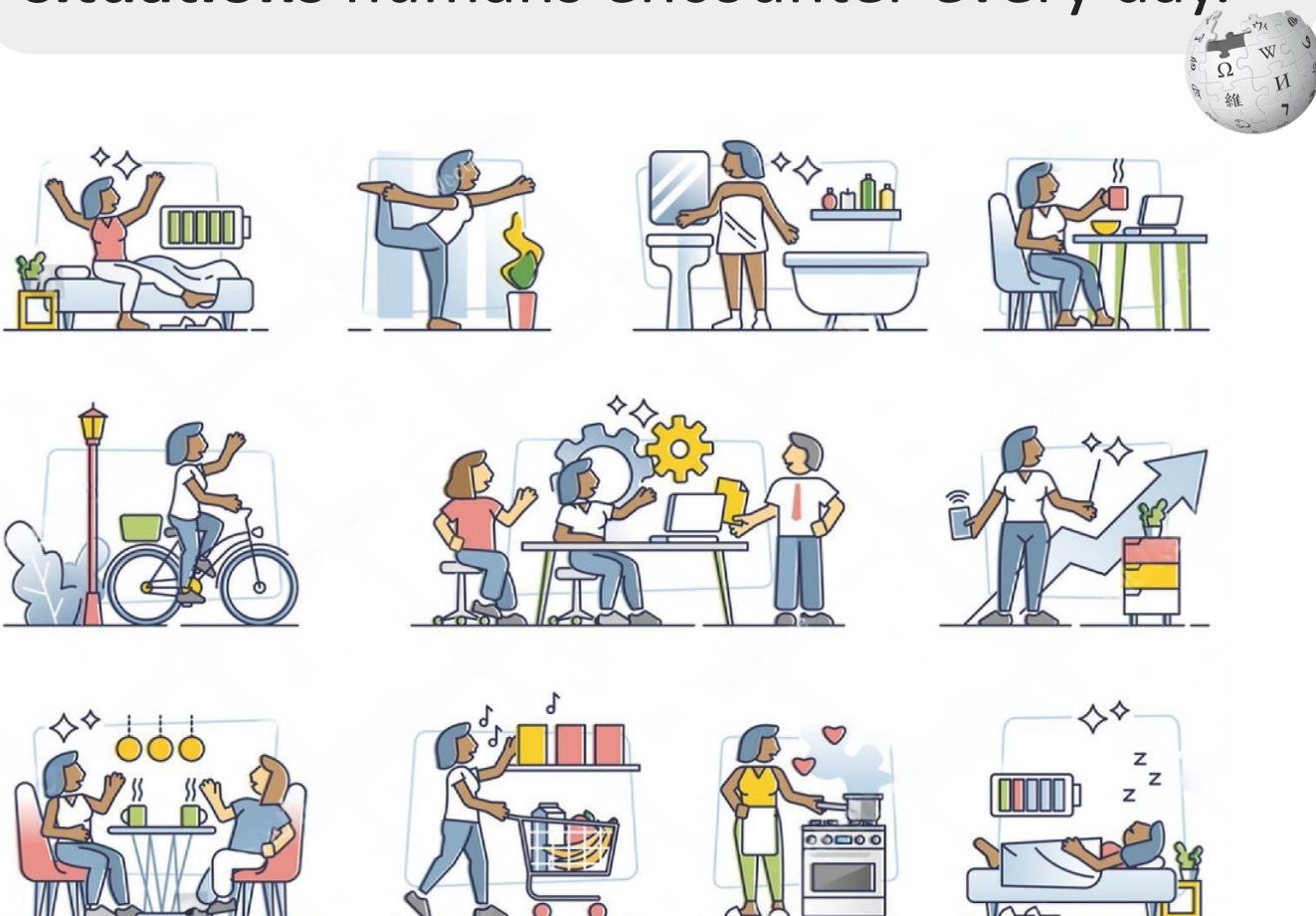
Question: What is the man holding in his hands?



Predicted top-5 answers with confidence:

✗ backpack	97.847%
bag	1.109%
✓ camera	0.541%
luggage	0.137%
strap	0.117%

**“ commonsense reasoning is a human-like ability to make presumptions about the type and essence of ordinary situations humans encounter every day. ”**



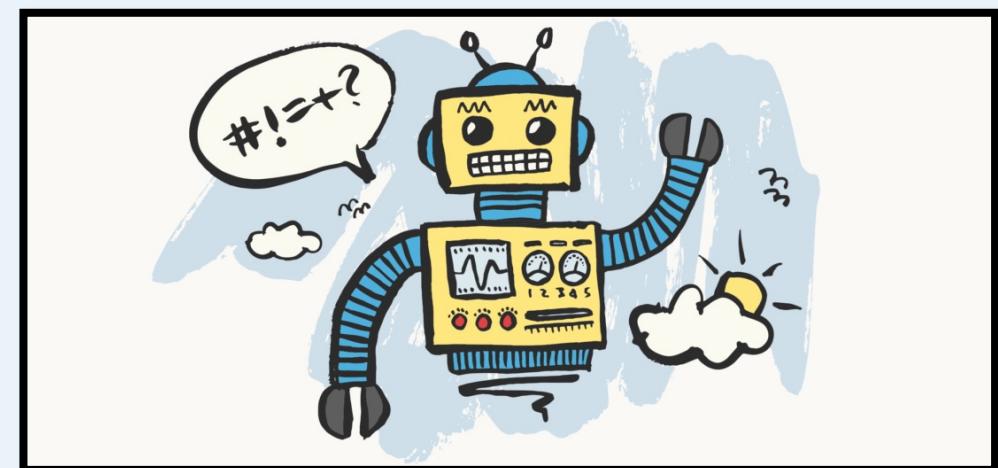
# Human-level AI !

Everyday  
Situations

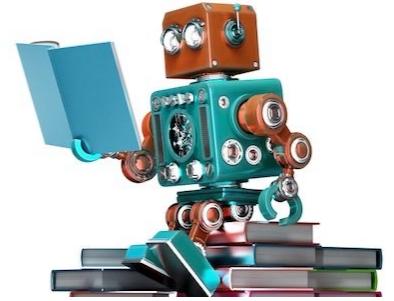
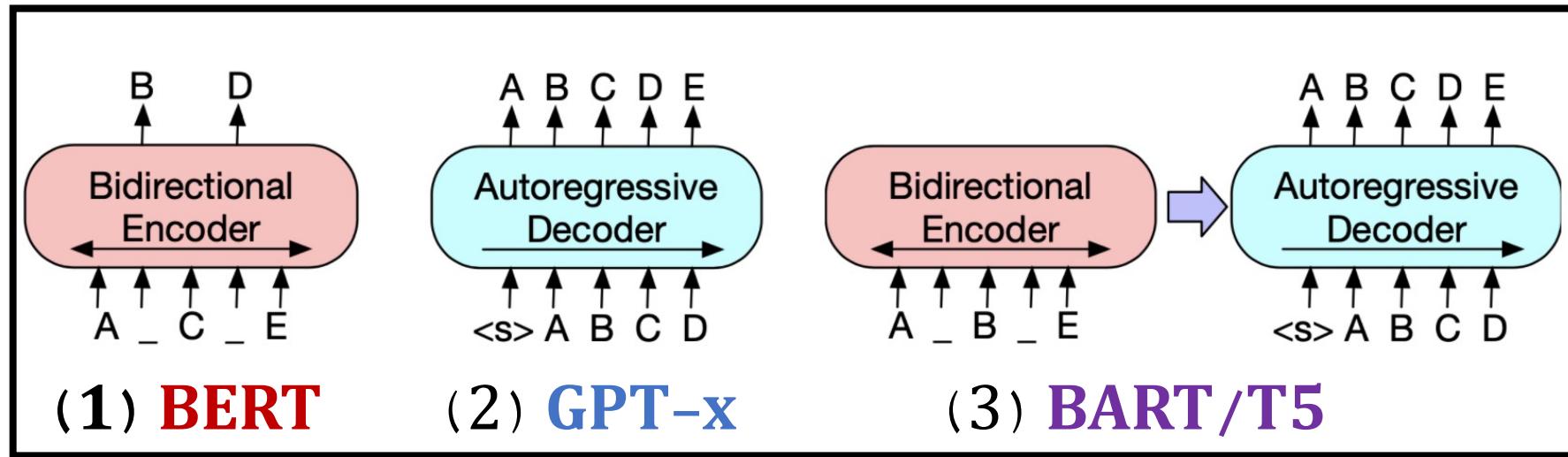


Think 🤔  
Talk 💬  
Act 💪

Common-Sense Reasoning



# Neural Language Models

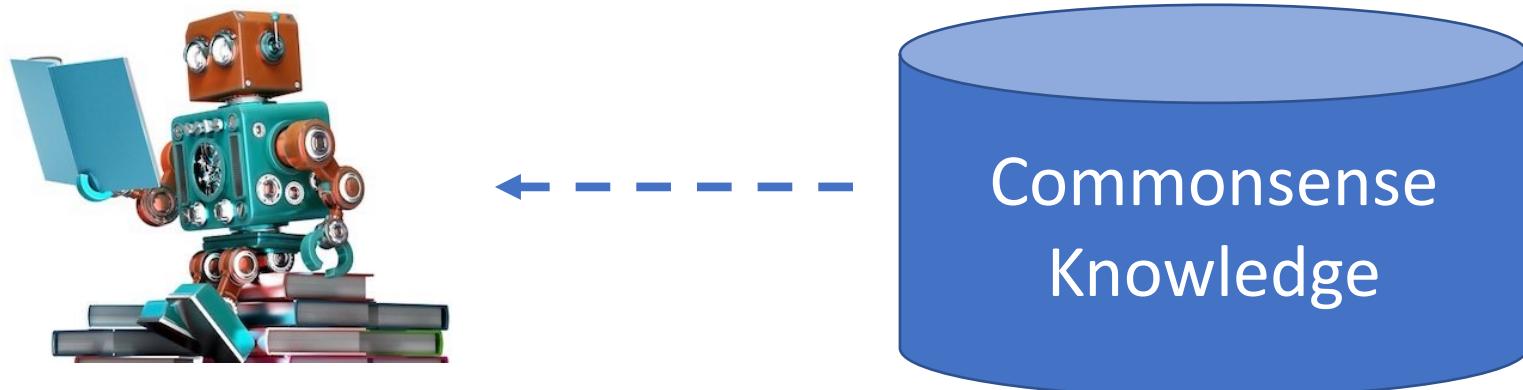
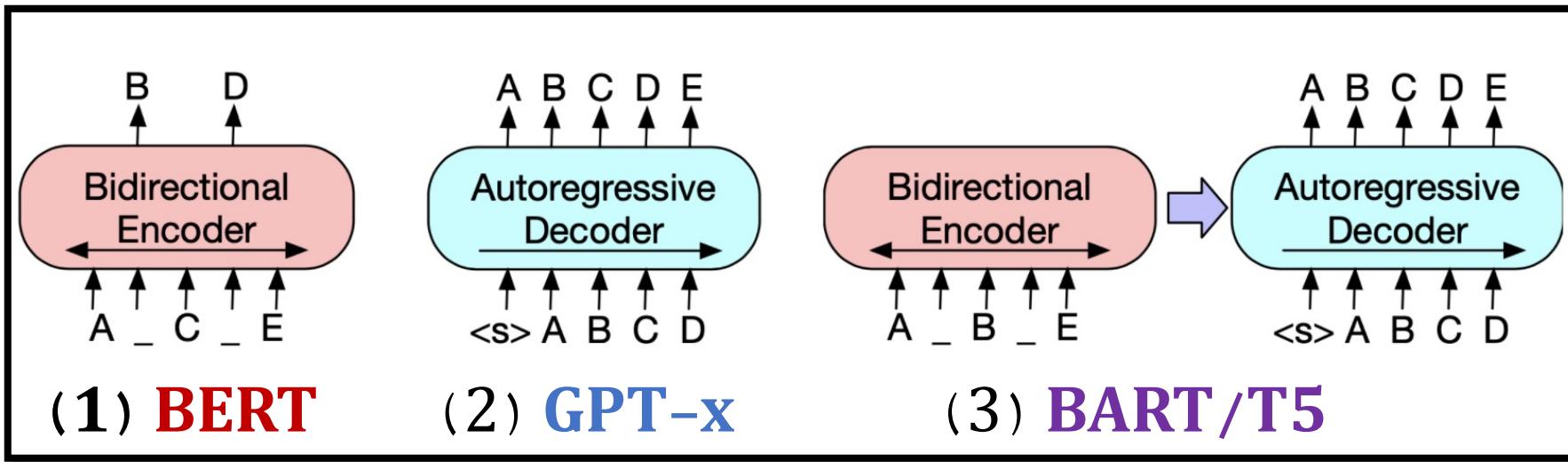


How do we incorporate commonsense knowledge into neural NLP models?

# Outline

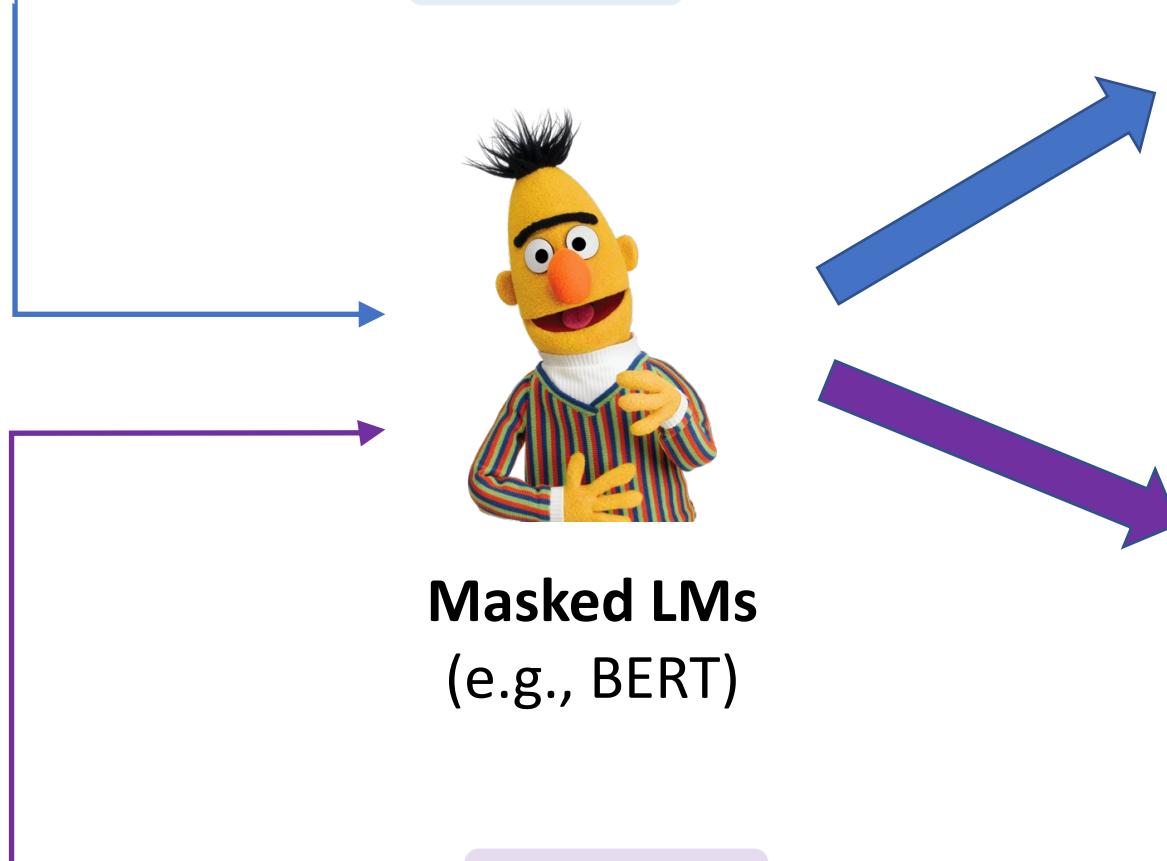
- → **Background of Commonsense Reasoning in NLP**
- **Incorporating Structured Commonsense Knowledge**
- **Incorporating Unstructured Commonsense Knowledge**
- **Incorporating Commonsense Knowledge for Generation**

# Neural Language Models



# Do LMs have common sense? (masked LMs)

Birds usually can [MASK].



Tigers usually have [MASK] legs.

LAMA (Petroni et al. 2019)

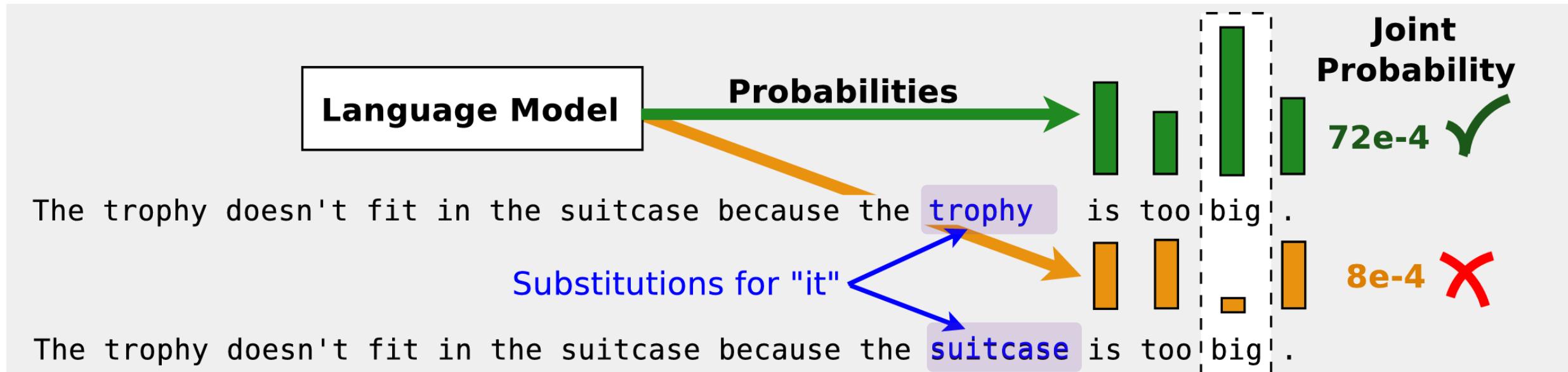
Prediction	Score	
Birds usually can fly .	33.1%	✓
Birds usually can sing .	8.2%	✓
Birds usually can survive .	3.5%	

NumerSense (Lin et al. 2020)

Prediction	Score	
Tigers usually have two legs .	14.1%	✗
Tigers usually have short legs .	11.2%	
Tigers usually have four legs .	8.8%	✓

# Do LMs have common sense? (auto-regressive LMs)

The **trophy** doesn't fit in the **suitcase** because **it** is too big.



*RNN-based LMs on WSC-237  
→ acc: 63.7% (random = 50%)*

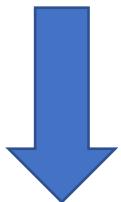
*Trinh & Le (2018)*

# Do LMs have common sense? (encoder-decoder LMs)

## Input:

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

- (A) fountain pen (B) pencil case (C) printer (D) notepad



**UNIFIED-QA** A *T5-based multi-task LM for QA*



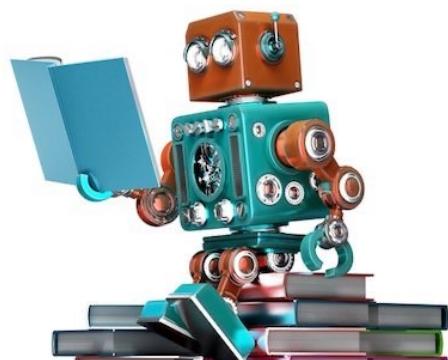
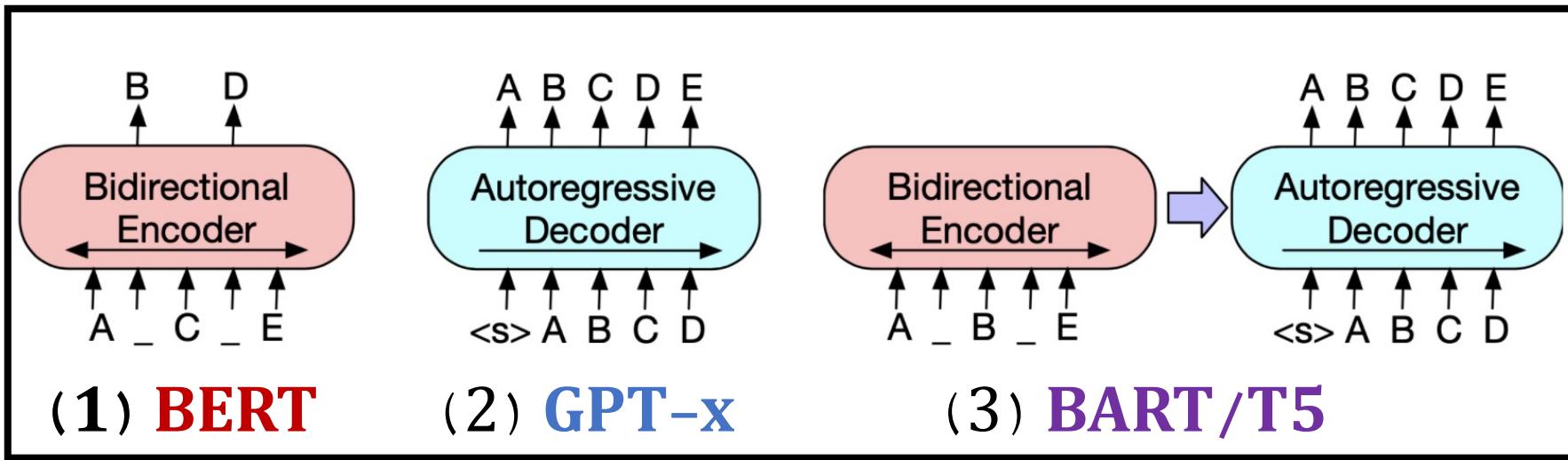
## Output:

Prediction [small, 60 million parameters]: pencil case

(Khashabi et al. 2020)

Prediction [large, 770 million parameters]: printer

# Neural Language Models



Better performance  
Interpretable  
Trustworthy



# MULTIPLE-CHOICE QA

What do you fill with ink to write notes on a piece of copy paper?  
(A) fountain pen (B) pencil case (C) printer (D) notepad



State-of-the-art QA Model

Prediction [small, 60 million parameters]: pencil case  
Prediction [large, 770 million parameters]: printer

**CommonsenseQA (Talmor et al. 2019)**

In the school play, Robin played a hero in the struggle to the death with the angry villain.

Q How would others feel afterwards?

- A (a) sorry for the villain  
(b) hopeful that Robin will succeed ✓  
(c) like Robin should lose

**Social IQA (Sap et al. 2019)**



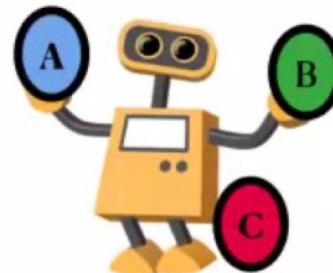
**Causal COPA**



**General CommonsenseQA WinoGrande**



**Common-Sense Benchmarks**



**Multiple-Choice Question Answering**

# OPEN-ENDED QA & TEXT GENERATION

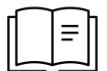
## OpenCSR (NAACL 2021)

Q: What can help alleviate global warming?



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.

## CommonGen (EMNLP 2020)

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. [Humans]
- 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.



GPT2: A dog throws a frisbee at a football player. [Machines]

UniLM: Two dogs are throwing frisbees at each other .

BART: A dog throws a frisbee and a dog catches it.

T5: dog catches a frisbee and throws it to a dog

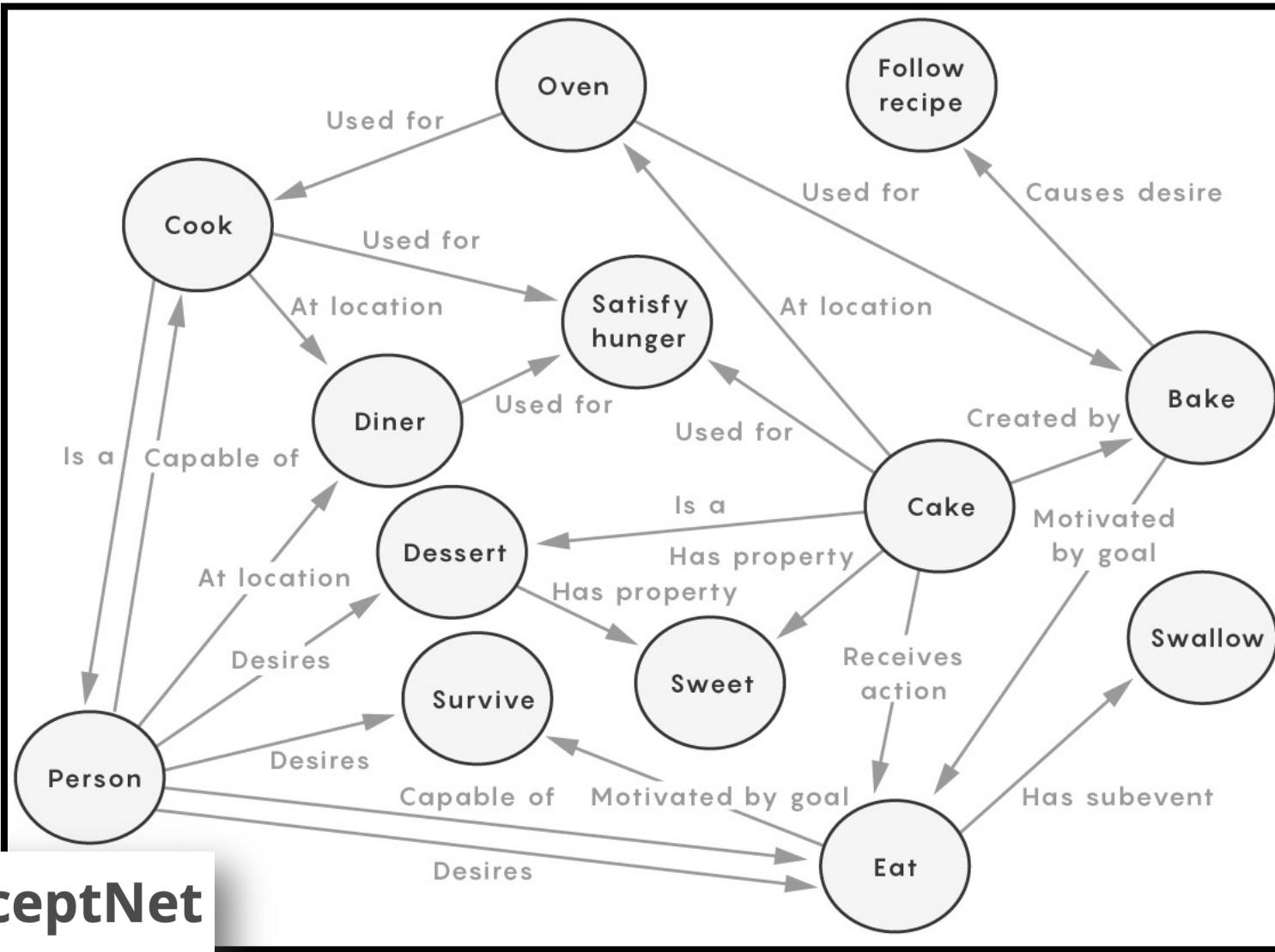


# Outline

- ~~Background of Commonsense Reasoning in NLP~~
- → **Incorporating Structured Commonsense Knowledge**
  - **KagNet (EMNLP 2019)**
  - **MHGRN (EMNLP 2020)**
  - **QA-GNN (NAACL 2021)**
  - GreaseLM (ICLR 2022)
  - GSC (ICLR 2022)
- Incorporating Unstructured Commonsense Knowledge
- Incorporating Commonsense Knowledge for Generation

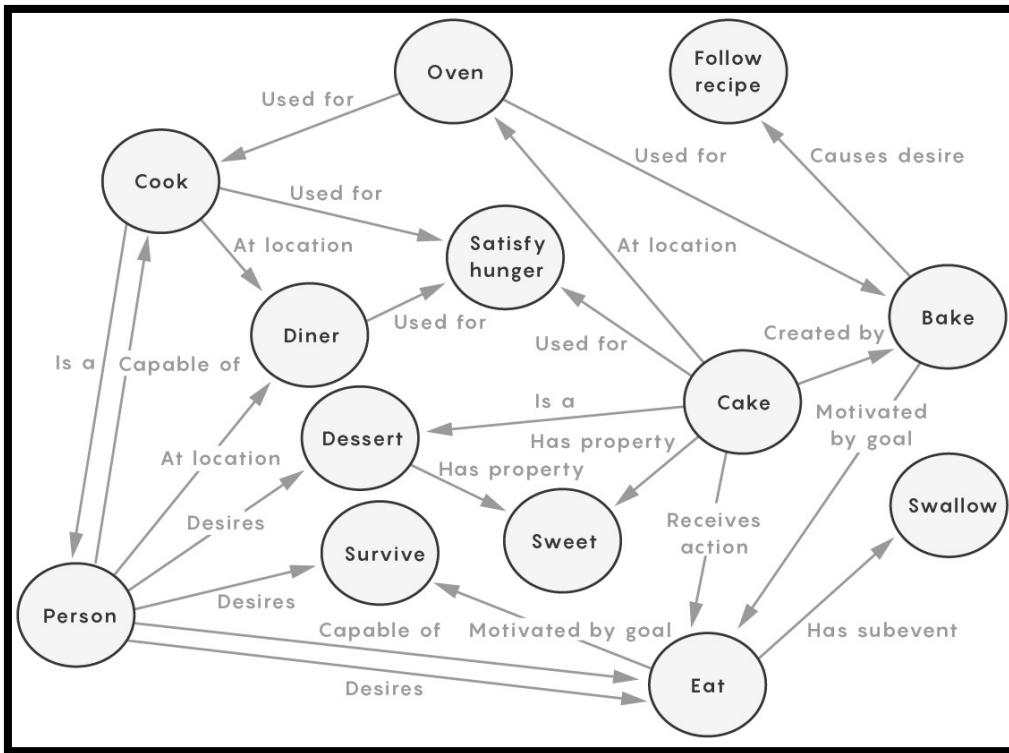


# STRUCTURED COMMONSENSE KNOWLEDGE BASE

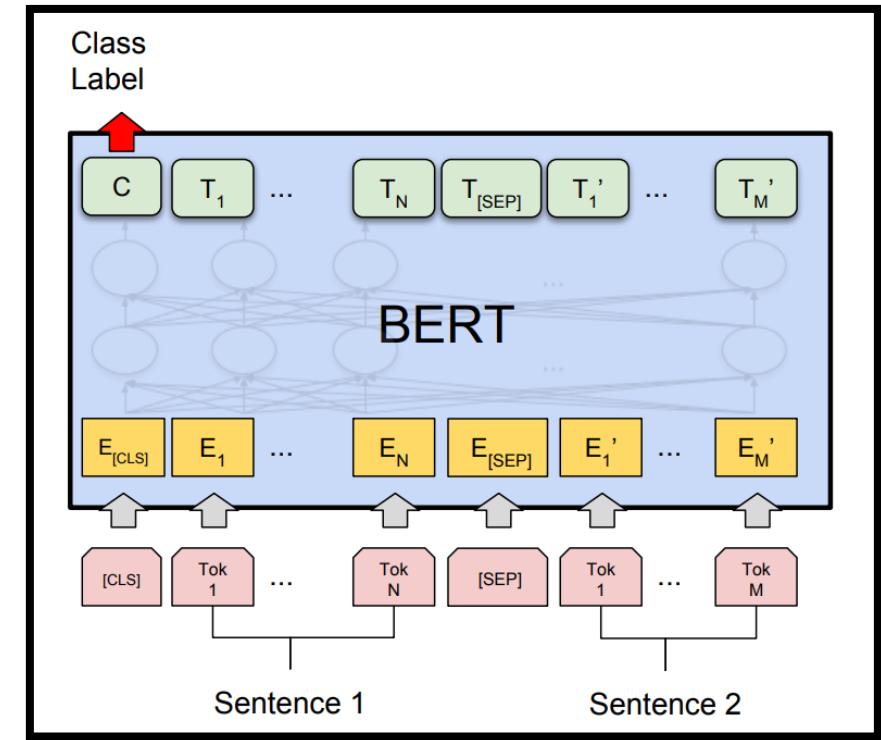


# STRUCTURED CSKB & NEURAL LMs

symbolic knowledge



neural networks



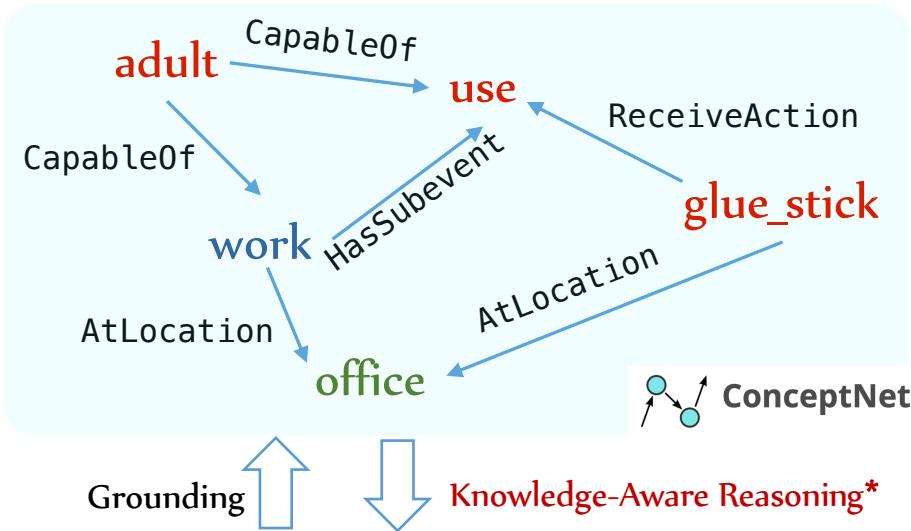
Structured Commonsense Knowledge Graphs  
(e.g., ConceptNet)

Pre-Trained Neural Language Models  
(e.g., BERT)

# Using Structured Commonsense KGs!

Symbol Space

Semantic Space



A Schema Graph  
for the choice B: office

Where do adults use glue sticks?

A: classroom B: office C: desk drawer

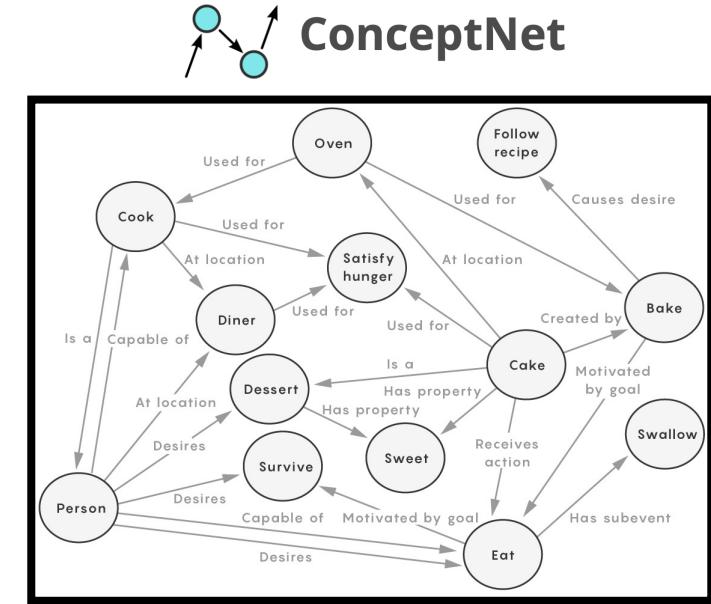
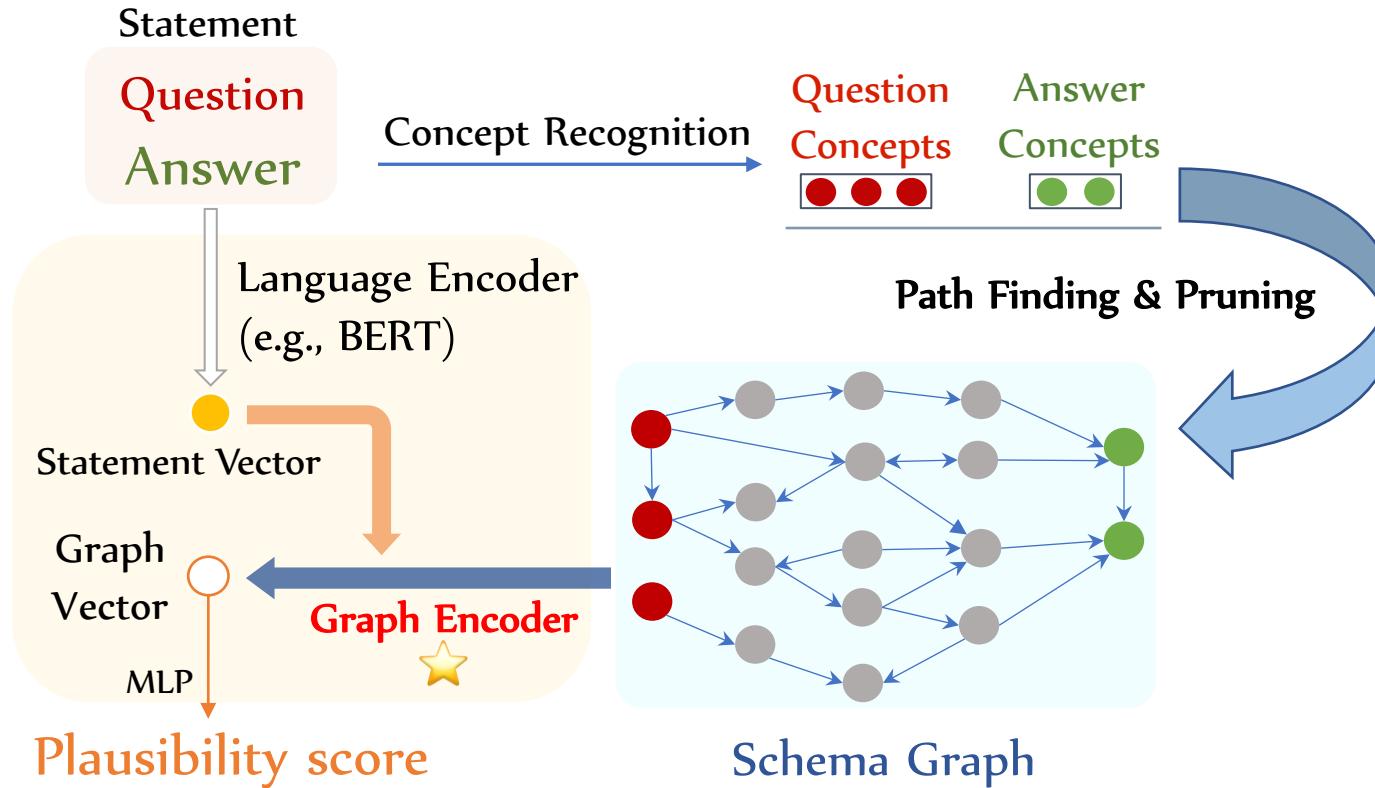
Question

Answer Options

Key Motivation:

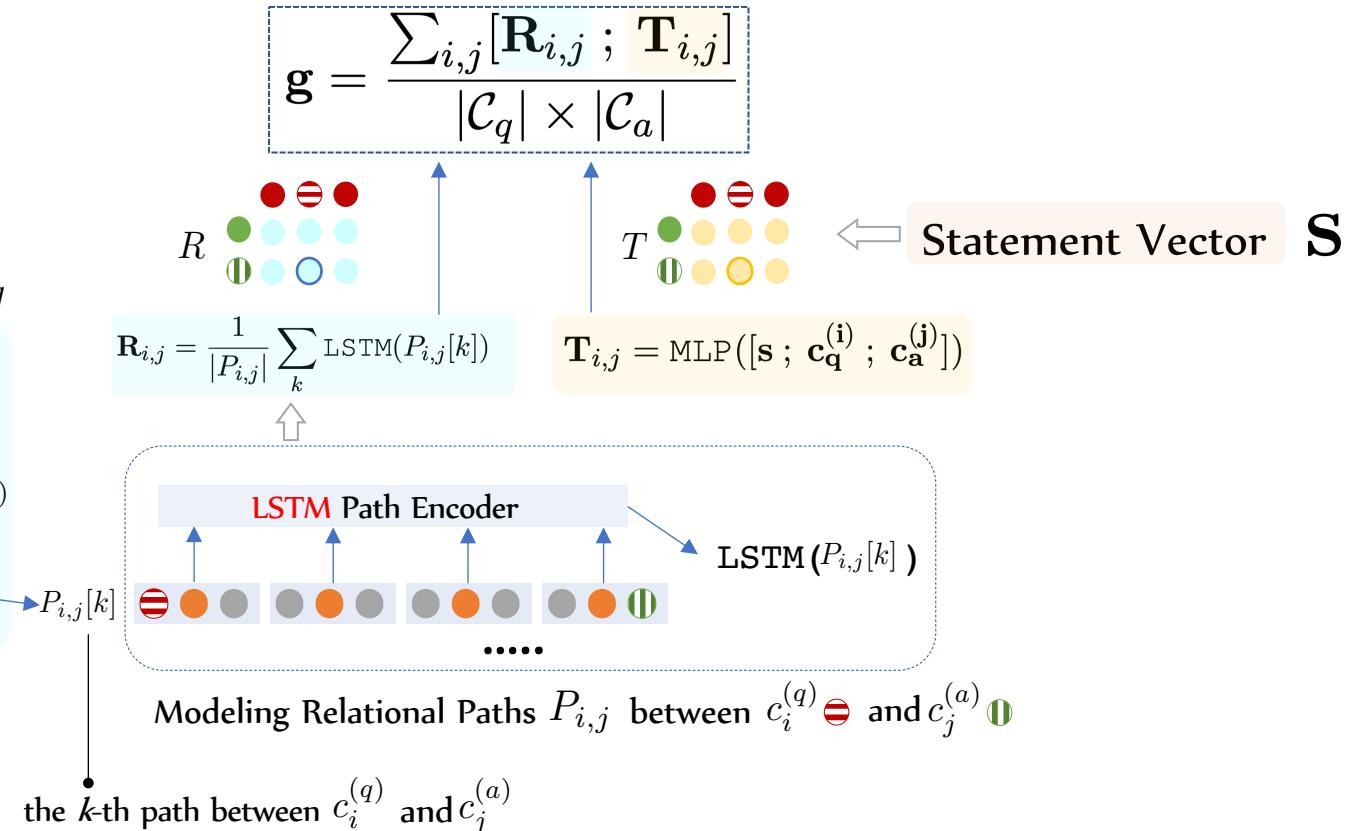
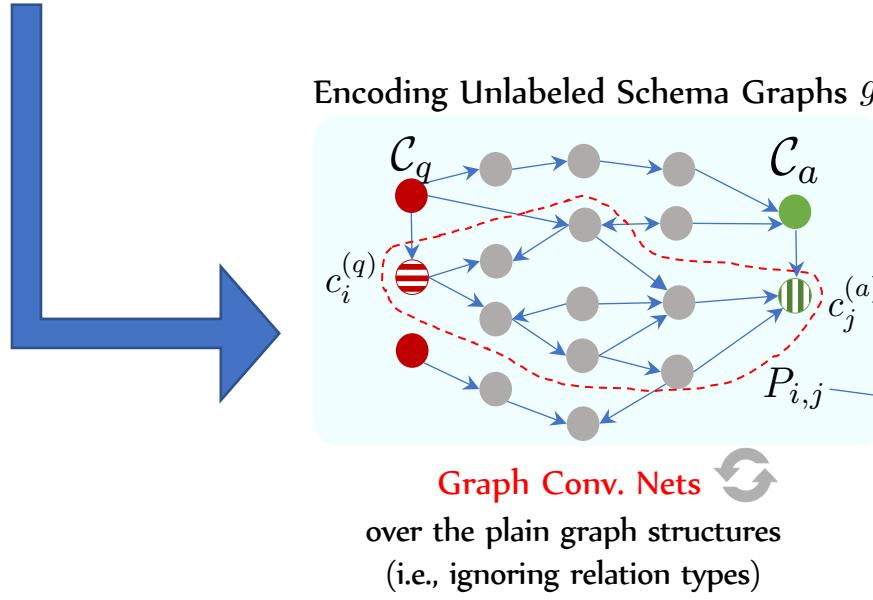
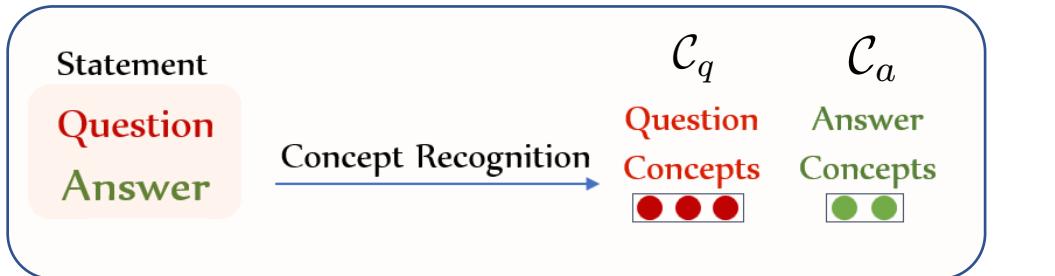
*using commonsense knowledge graphs (CSKGs) to support language reasoning!*

# KagNet: Knowledge-Aware Graph Neural Networks



Knowledge Graph-Augmented Language Reasoning

# KagNet: Knowledge-Aware Graph Neural Networks



**Path-based Relational Graph Encoder + Language Representation**

# KagNet: Knowledge-Aware Graph Neural Networks

- Two average pooling:

- Assuming all QA-concept pairs are equally important  $\mathbf{g} = \frac{\sum_{i,j} [\mathbf{R}_{i,j} ; \mathbf{T}_{i,j}]}{|\mathcal{C}_q| \times |\mathcal{C}_a|}$
- Assuming all paths are equally relevant  $\mathbf{R}_{i,j} = \frac{1}{|P_{i,j}|} \sum_k \text{LSTM}(P_{i,j}[k])$

- Modeling the two-level importance as latent weights:

$$\alpha_{(i,j,k)} = \mathbf{T}_{i,j} \mathbf{W}_1 \text{LSTM}(P_{i,j}[k]),$$

$$\hat{\alpha}_{(i,j,\cdot)} = \text{SoftMax}(\alpha_{(i,j,\cdot)}),$$

$$\hat{\mathbf{R}}_{i,j} = \sum_k \hat{\alpha}_{(i,j,k)} \cdot \text{LSTM}(P_{i,j}[k])$$

Path-Level Attention  
(attending on semantic space)

$$\beta_{(i,j)} = \mathbf{s} \mathbf{W}_2 \mathbf{T}_{i,j}$$

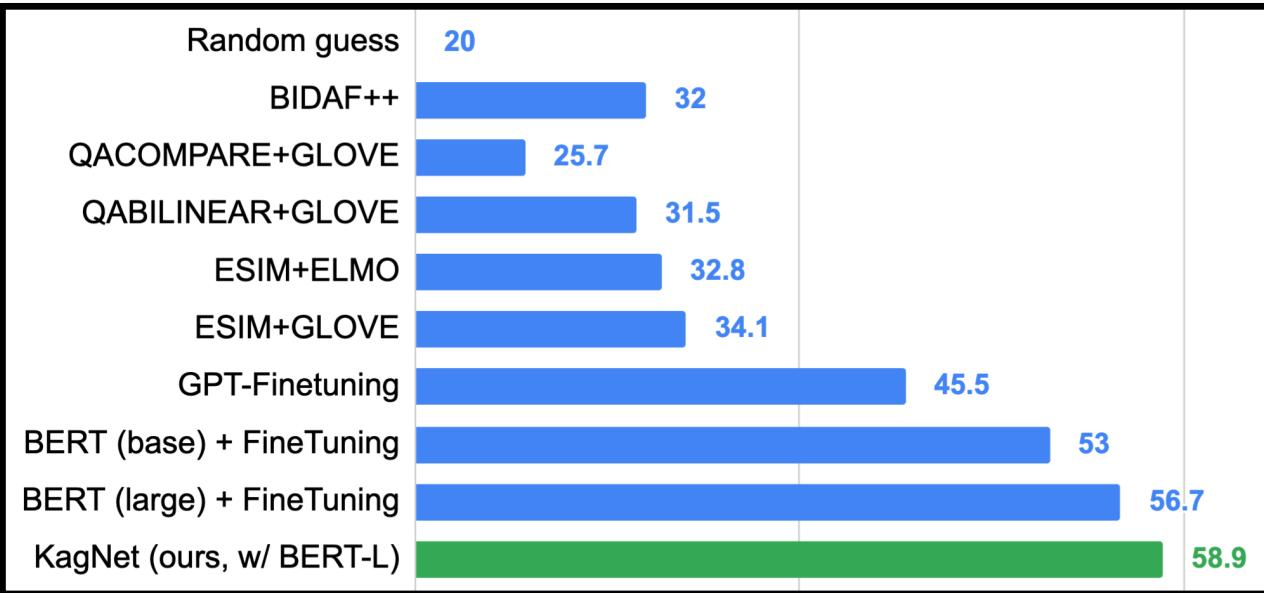
$$\hat{\beta}_{(\cdot,\cdot)} = \text{SoftMax}(\beta_{(\cdot,\cdot)})$$

$$\hat{\mathbf{g}} = \sum_{i,j} \hat{\beta}_{(i,j)} [\hat{\mathbf{R}}_{i,j} ; \mathbf{T}_{i,j}]$$

Pair-Level Attention  
(attending on statement)

+ Hierarchical Path-based Attention

# Experimental Results



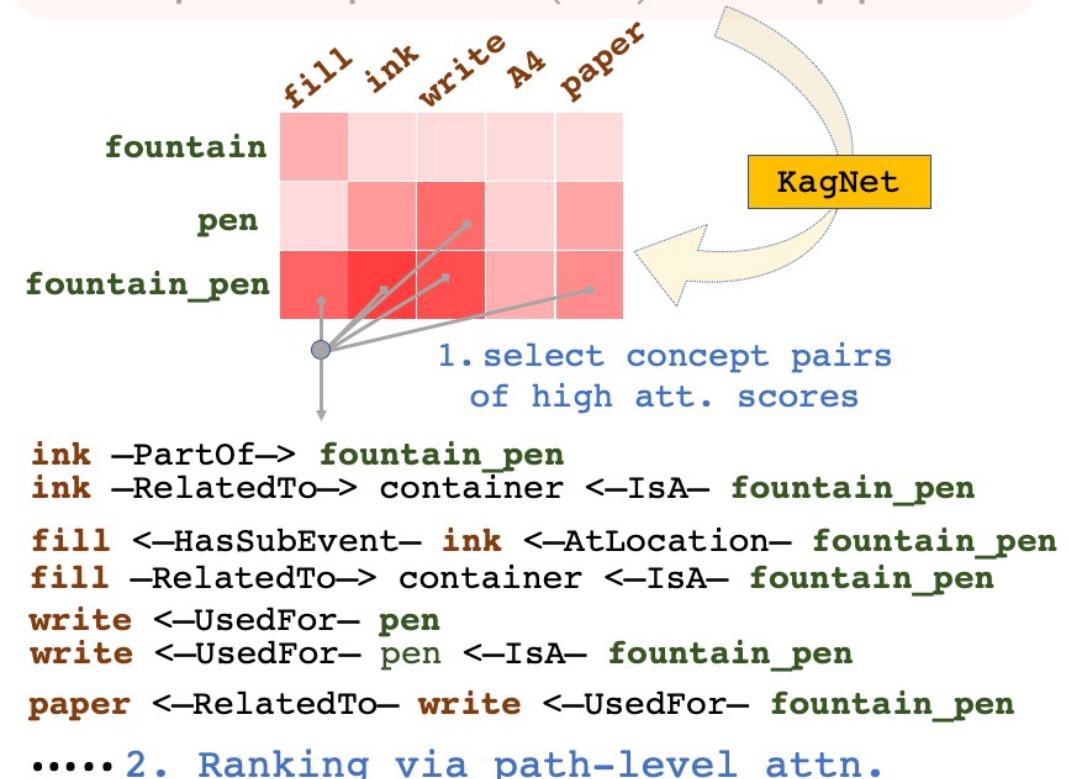
(as of the publication time)

The latest results on the official leaderboard:

<https://www.tau-nlp.org/csqa-leaderboard>.

What do you **fill** with **ink** to **write** on an A4 paper?

- A: fountain pen ✓ (KagNet); B: printer (BERT);  
C: squid D: pencil case (GPT); E: newspaper

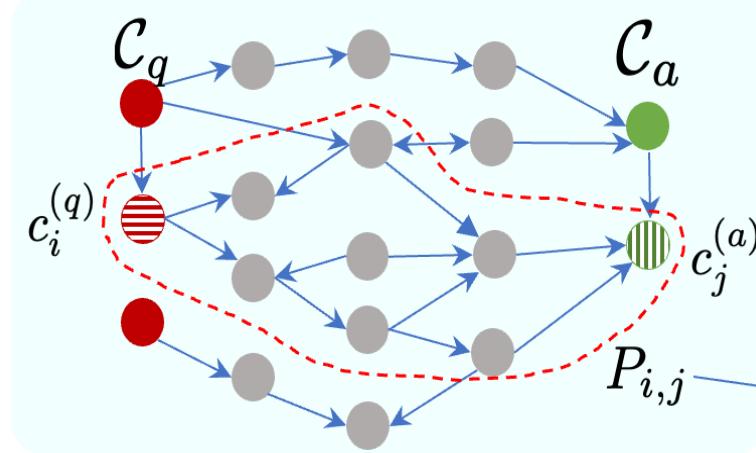


An **interpretable**  
reasoning process!

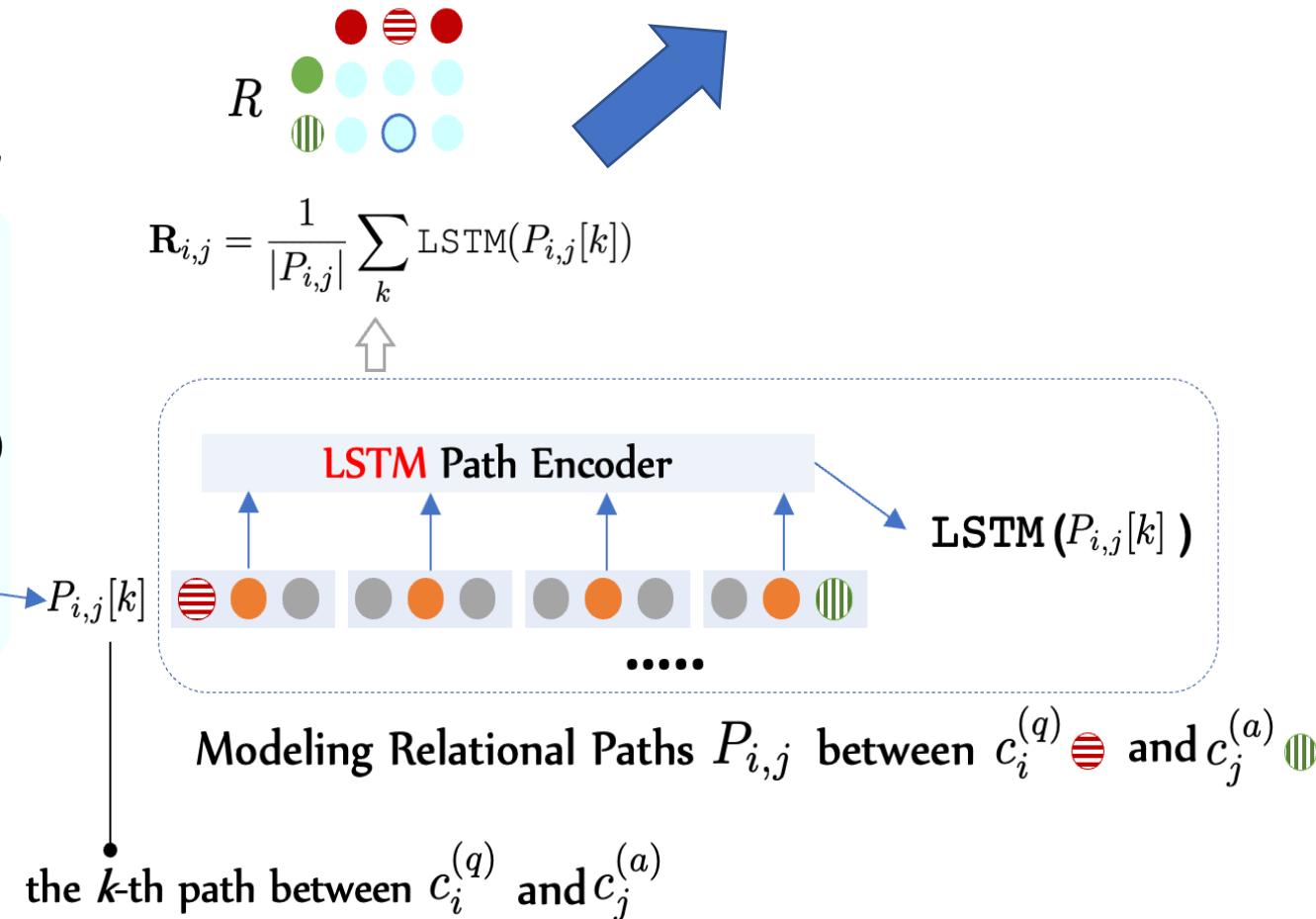
# KagNet: The Scalability Issue

- ⚠ Larger number of paths.
- ⚠ Longer sequence of triples.

Encoding Unlabeled Schema Graphs  $g$



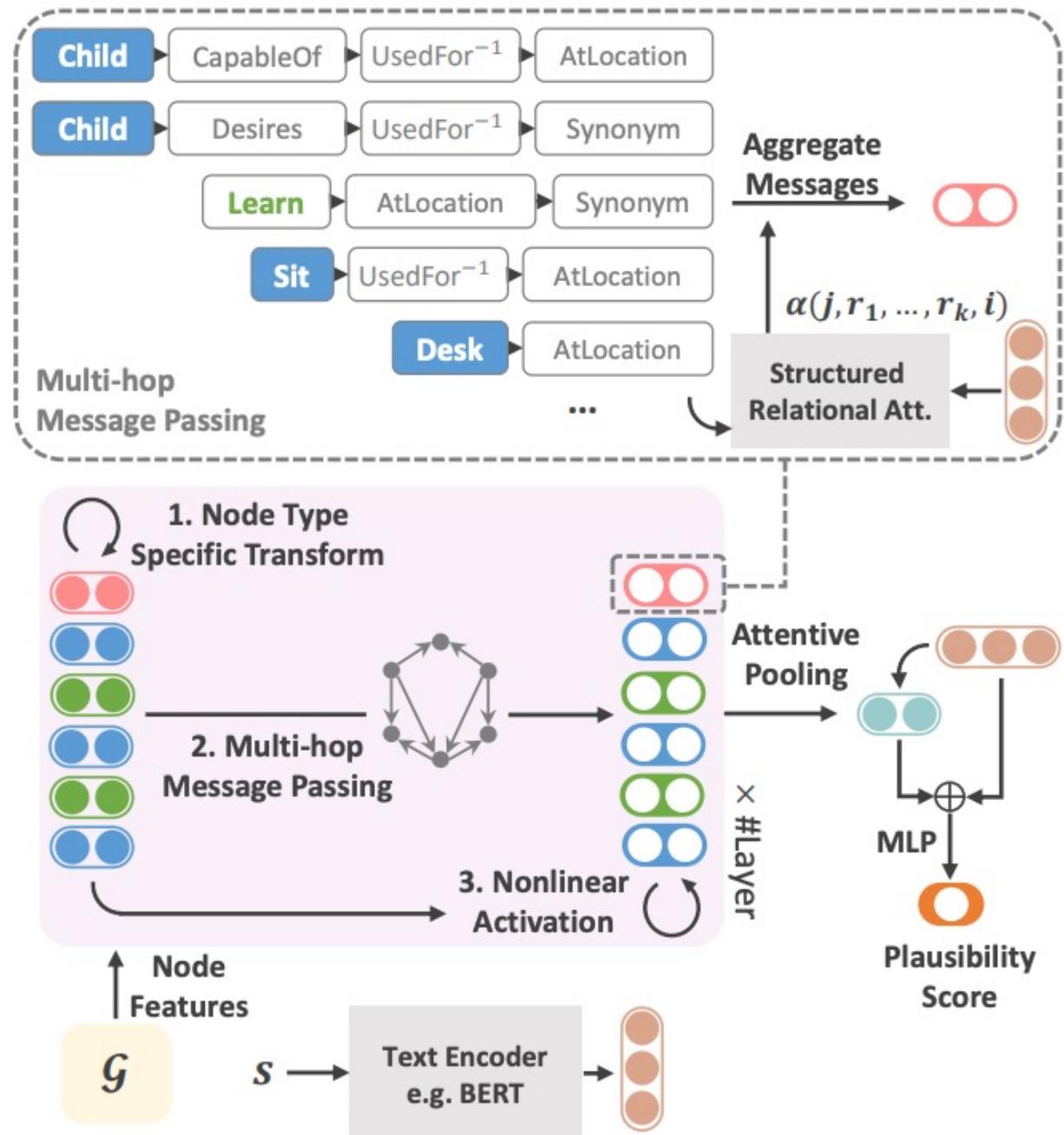
Graph Conv. Nets   
over the plain graph structures  
(i.e., ignoring relation types)



# MHGRN (EMNLP 20)

Can we **directly** model **multi-hop message passing** in relational graph networks?

Relational paths can be modeled as **multi-hop message passing** with **multi-layer graph attention networks**.



$$\text{GNN}(\mathcal{G}) = \text{Pool}(\{\mathbf{h}_1', \mathbf{h}_2', \dots, \mathbf{h}_n'\}).$$

$$\mathbf{h}_i' = \sigma \left( \left( \sum_{r \in \mathcal{R}} |\mathcal{N}_i^r| \right)^{-1} \sum_{r \in \mathcal{R}} \sum_{j \in \mathcal{N}_i^r} \mathbf{W}_r \mathbf{h}_j \right)$$

**Relational GNN (e.g., R-GCN)**

**Simple, fast**

**Bad representation of k-hop**

**Normalized Attention**

**Path-based, with attention**  
**Slow, cannot scale**

$$\text{KAGNET}(\mathcal{G}) = \text{Pool} \left( \{\text{LSTM}(j, r_1, \dots, r_k, i) \mid (j, r_1, j_1), \dots, (j_{k-1}, r_k, i) \in \mathcal{E}, 1 \leq k \leq K\} \right)$$

**KagNet (paths are pre-filtered)**

## Multi-Hop Graph Relation Nets

$$\Phi_k = \{(j, r_1, \dots, r_k, i) \mid (j, r_1, j_1), \dots, (j_{k-1}, r_k, i) \in \mathcal{E}\} \quad (1 \leq k \leq K).$$

$$\begin{aligned} \mathbf{z}_i^k &= \sum_{(j, r_1, \dots, r_k, i) \in \Phi_k} \alpha(j, r_1, \dots, r_k, i) / d_i^k \cdot \mathbf{W}_0^K \\ &\cdots \mathbf{W}_0^{k+1} \mathbf{W}_{r_k}^k \cdots \mathbf{W}_{r_1}^1 \mathbf{x}_j \quad (1 \leq k \leq K) \\ &\mathbf{W}_r^t \quad (1 \leq t \leq K, 0 \leq r \leq m) \\ &\text{low-rank approximation} \end{aligned}$$

$$\begin{aligned} \mathbf{z}_i &= \sum_{k=1}^K \text{softmax}(\text{bilinear}(\mathbf{s}, \mathbf{z}_i^k)) \cdot \mathbf{z}_i^k \\ \mathbf{h}_i' &= \sigma(\mathbf{V} \mathbf{h}_i + \mathbf{V}' \mathbf{z}_i) \end{aligned}$$

Model	Time	Space
<i>G is a dense graph</i>		
K-hop KagNet	$\mathcal{O}(m^K n^{K+1} K)$	$\mathcal{O}(m^K n^{K+1} K)$
K-layer RGCN	$\mathcal{O}(mn^2 K)$	$\mathcal{O}(mnK)$
MHGRN	$\mathcal{O}(m^2 n^2 K)$	$\mathcal{O}(mnK)$
<i>G is a sparse graph with maximum node degree <math>\Delta \ll n</math></i>		
K-hop KagNet	$\mathcal{O}(m^K n K \Delta^K)$	$\mathcal{O}(m^K n K \Delta^K)$
K-layer RGCN	$\mathcal{O}(mn K \Delta)$	$\mathcal{O}(mnK)$
MHGRN	$\mathcal{O}(m^2 n K \Delta)$	$\mathcal{O}(mnK)$

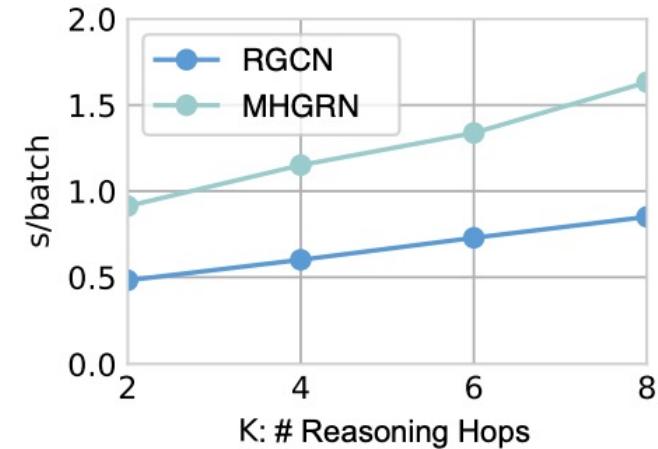
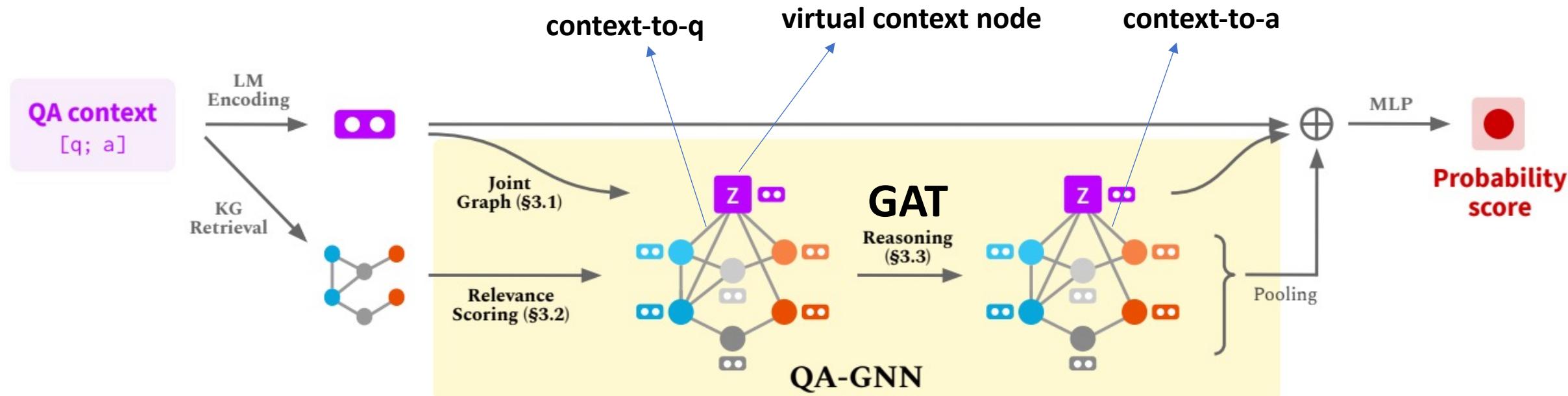
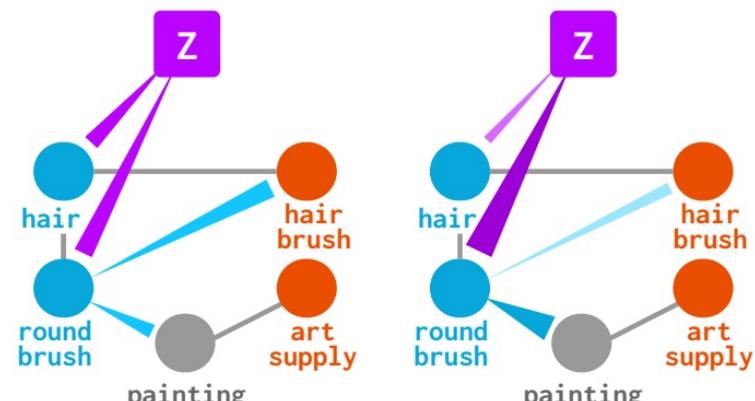


Figure 7: **Analysis of model scalability.** Comparison of per-batch training efficiency w.r.t. # hops  $K$ .

# QA-GNN (NAACL 2021)



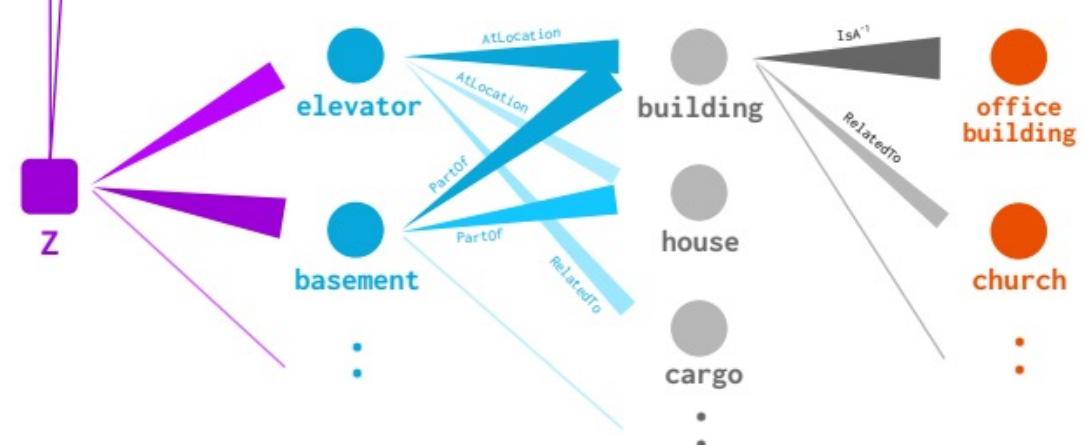
If it is **not** used for **hair**, a **round brush** is an example of what?  
 A. hair brush B. art supply\*



- A. hair brush (0.38)  
**B. art supply (0.64)**

(a) Attention visualization direction: BFS from **Q**

Where would you find a **basement** that can be accessed with an **elevator**?  
 A. closet B. church C. office building\*



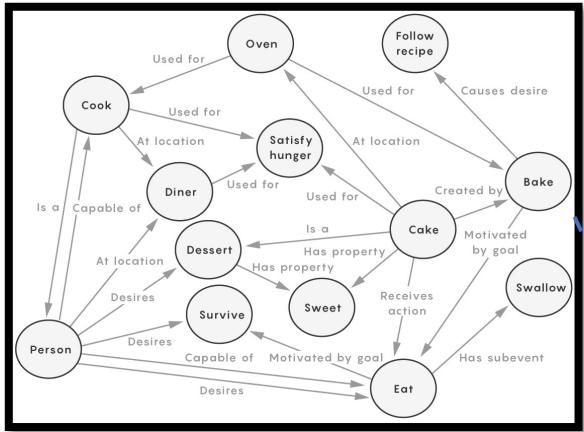
# KagNet vs MHGRN vs QA-GNN

Methods	IHdev-Acc. (%)	IHtest-Acc. (%)
RoBERTa-large (w/o KG)	73.07 ( $\pm 0.45$ )	68.69 ( $\pm 0.56$ )
+ RGCN (Schlichtkrull et al., 2018)	72.69 ( $\pm 0.19$ )	68.41 ( $\pm 0.66$ )
+ GconAttn (Wang et al., 2019a)	72.61 ( $\pm 0.39$ )	68.59 ( $\pm 0.96$ )
+ KagNet (Lin et al., 2019)	73.47 ( $\pm 0.22$ )	69.01 ( $\pm 0.76$ )
+ RN (Santoro et al., 2017)	74.57 ( $\pm 0.91$ )	69.08 ( $\pm 0.21$ )
+ MHGRN (Feng et al., 2020)	74.45 ( $\pm 0.10$ )	71.11 ( $\pm 0.81$ )
+ QA-GNN (Ours)	<b>76.54</b> ( $\pm 0.21$ )	<b>73.41</b> ( $\pm 0.92$ )

Table 2: **Performance comparison on Commonsense QA in-house split** (controlled experiments). As the official test is hidden, here we report the in-house Dev (IHdev) and Test (IHtest) accuracy, following the data split of Lin et al. (2019).

# Talk Outline

- ~~Background of Commonsense Reasoning in NLP~~
- ~~Incorporating Structured Commonsense Knowledge~~
- → **Incorporating Unstructured Commonsense Knowledge**
  - DrFact (NAACL 2021)
- Incorporating Commonsense Knowledge for Generation



**Structured**  
**Commonsense KGs**  
(e.g., ConceptNet)

- **binary**
- **limited**
- **noisy & incomplete**

## Unstructured Commonsense Knowledge Base (e.g., GenericsKB)

Trees are perennial plants that have long woody trunks.

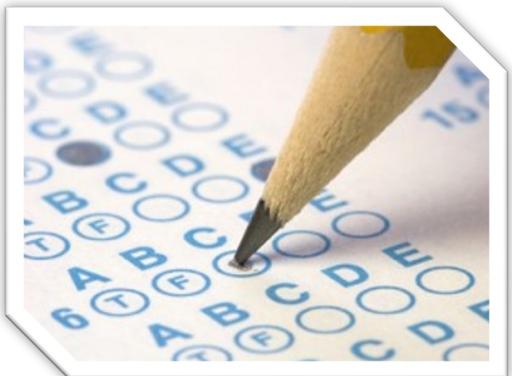


Most trees add one new ring for each year of growth.

Trees produce oxygen by absorbing carbon dioxide from the air.

Trees are large, generally single-stemmed, woody plants.

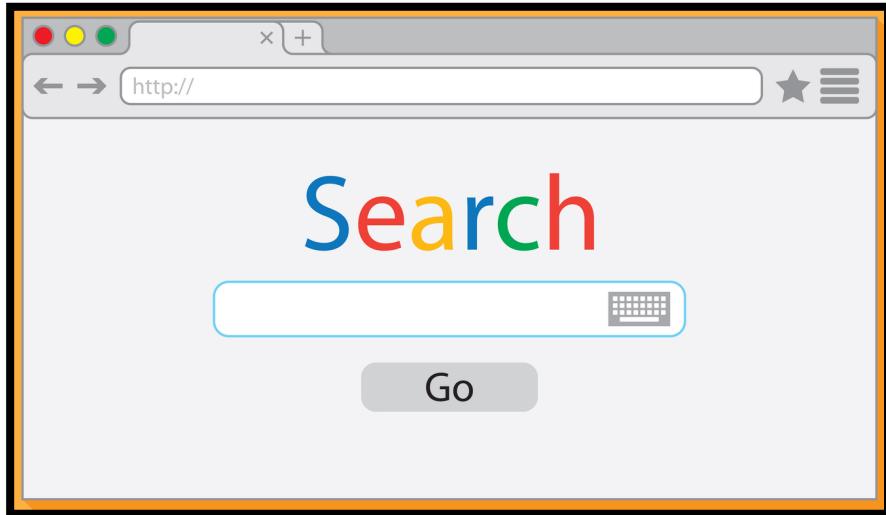
Trees grow using photosynthesis, absorbing carbon dioxide and releasing oxygen.



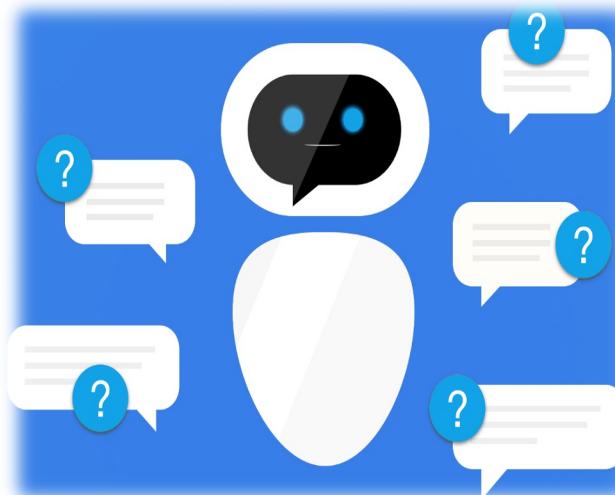
Open-Ended Question Answering  
(i.e., no choices provided)

# Is multiple-choice QA useful in realistic applications?

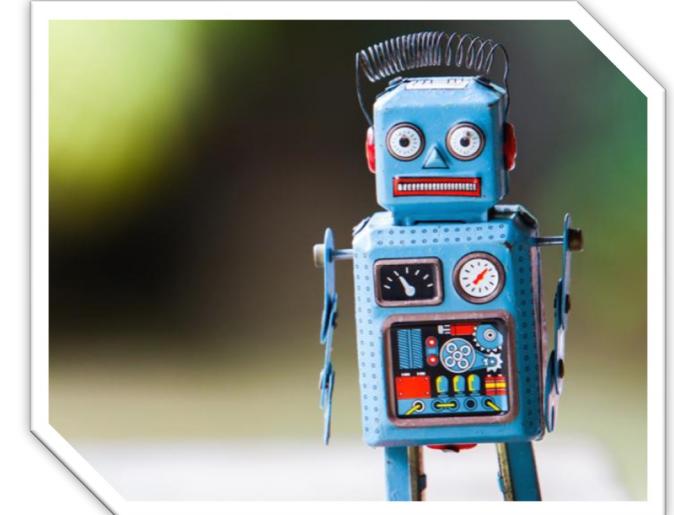
Search Engines



Chatbots



Robots



Common Sense as a Service for practical AI applications.

Real users usually do **NOT** have any *answer candidates* when querying commonsense knowledge.

# Open-Ended Commonsense Reasoning

Q: What can help alleviate global warming?



## 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.



## Multiple-Choice/Closed CSR

Input: a question + a few choices

- A) air conditioner B) fossil fuel
- C) **renewable energy** D) carbon dioxide



*Can machines learn to reason without answer candidates?*

# 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**.



$q_1 = \text{the dog in the TV show Family Guy}$   
↓  
 $q_2 = \text{who voices } [q_1. \text{ answer}]$

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

What can help alleviate global warming?



$q_1 = \text{what contributes to global warming}$   
↓  
 $q_2 = \text{what removes } [q_1. \text{ answer}]$

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).

# Notations for DrFact

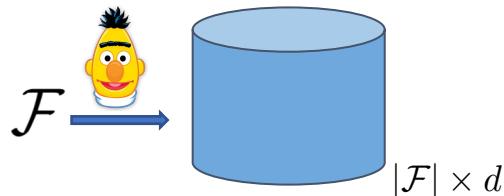
a **corpus** of common-sense facts, e.g., **GenericsKB**.  $\mathcal{F}$

$f_i \in \mathcal{F}$

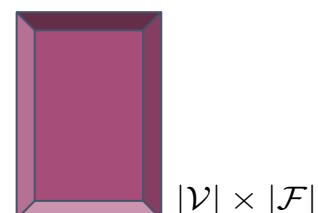
A **fact** is a sentence of generic commonsense knowledge

$c_j \in \mathcal{V}$

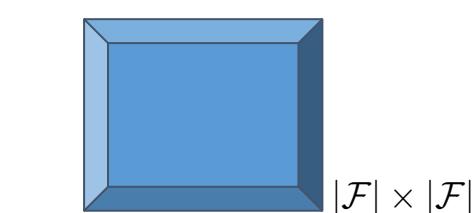
A **concept** is a noun or noun-chunk that are mentioned in  $\mathcal{F}$



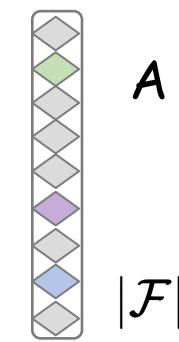
Dense Matrix  
of Fact Embeddings



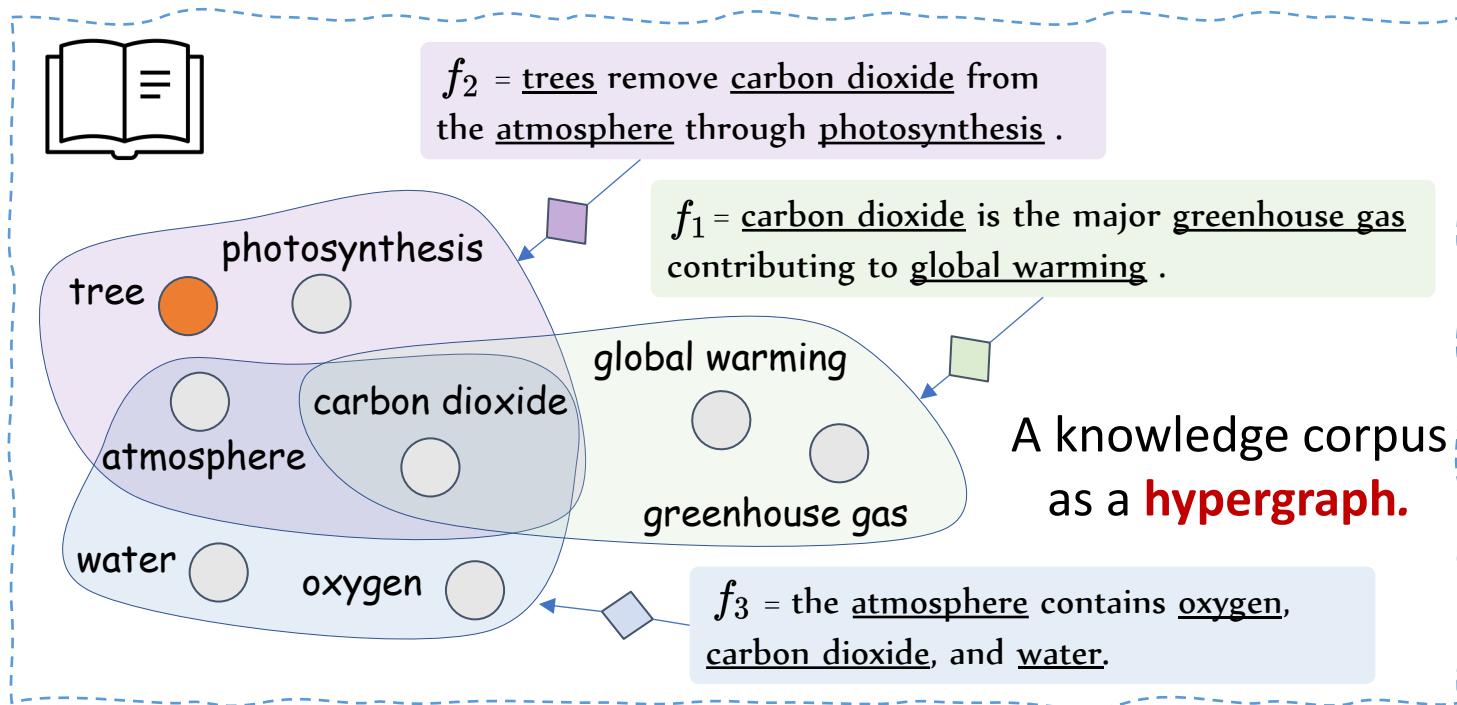
Sparse Matrix  
of Concept-to-Fact Links



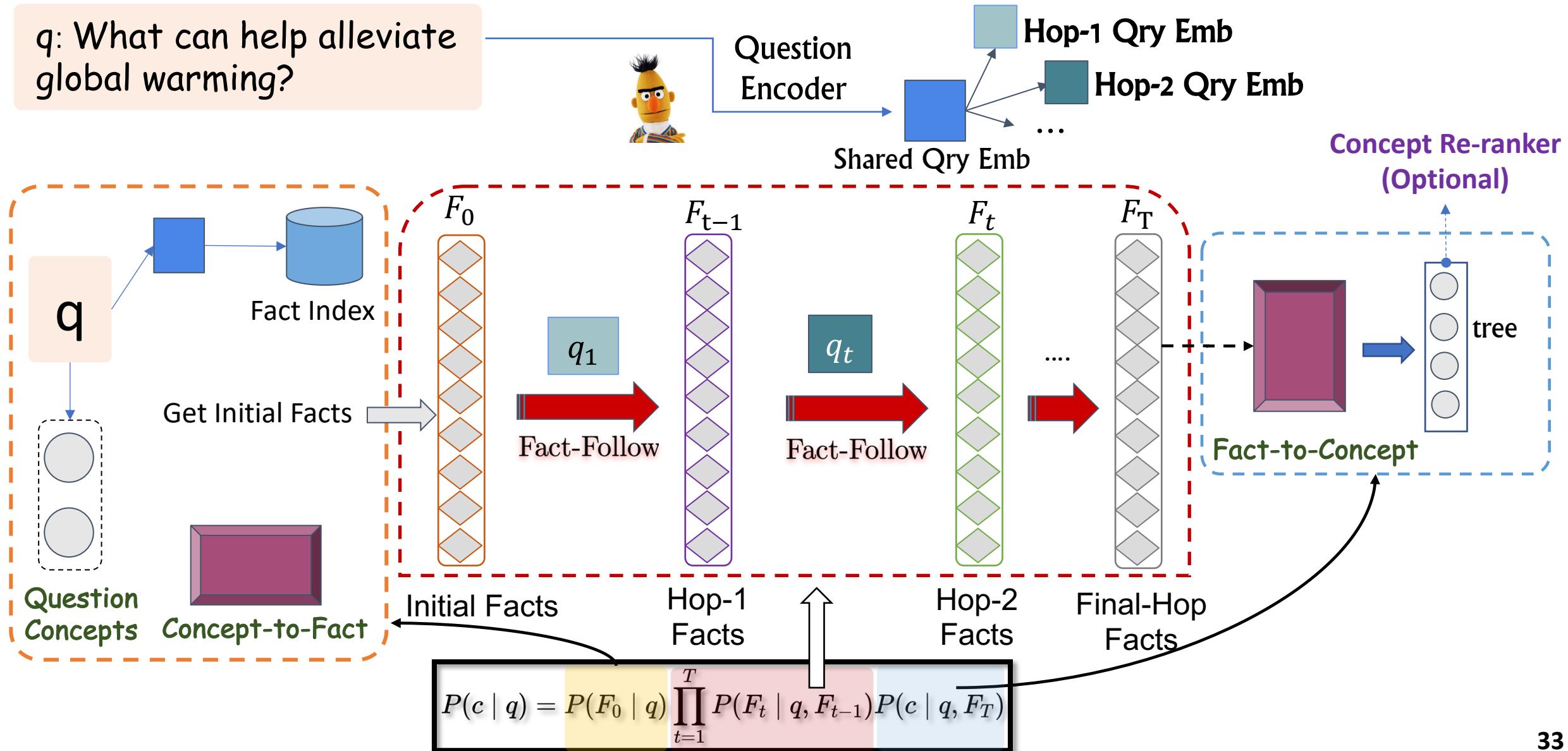
Sparse Matrix  
of Fact-to-Fact Links



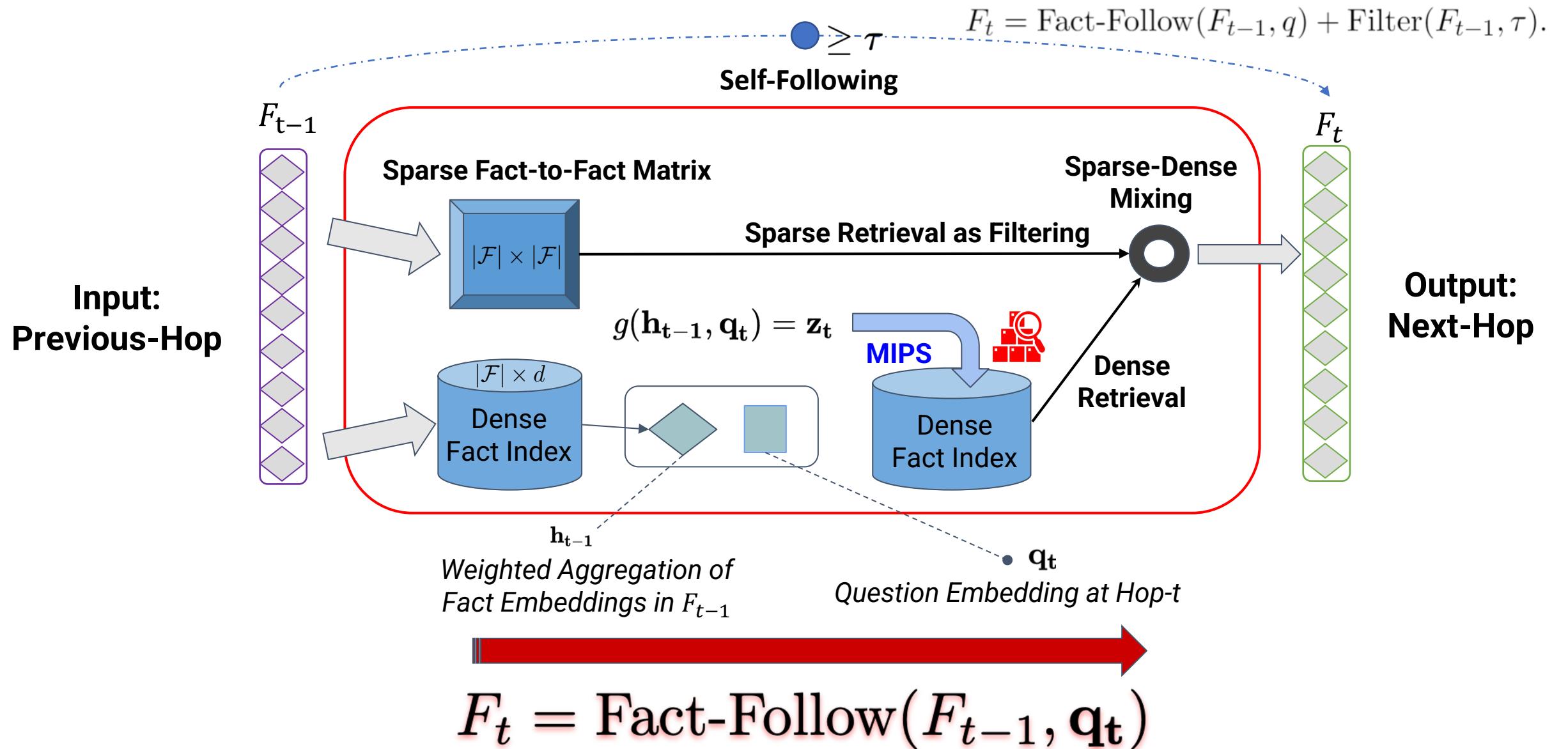
A weighted set of facts  
→  
A sparse vector.



# Overall Workflow of DrFact



# DrFact: Differentiable Fact-Following Operation



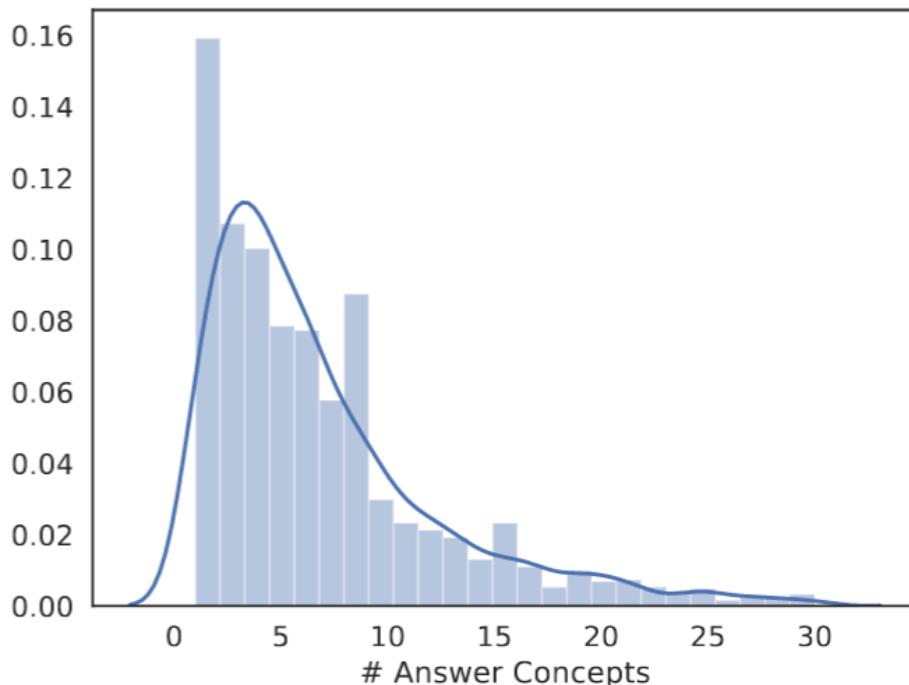
# Comparisons with baseline methods.

Methods	BM25 (off-the-shelf)	DPR (EMNLP 2020)	DrKIT (ICLR 2020)	DrFact (NAACL 2021)
Knowledge Structure	A set of documents	A set of documents	Mention-Entity Bipartite Graph	Concept-Fact Hypergraph
Multi-hop Reasoning Formulation	N/A	N/A	Entity-Following	Fact-Following
Index for Dense Retrieval	N/A	Dense Fact Embeddings	Dense Mention Embedding	Dense Fact Embeddings
Sparse Retrieval Method	TF-IDF based Index+ BM25 Ranking Func.	N/A	Entity Cooccurrence	Fact-to-Fact Matrix
Multi-Hop Questions	N/A	N/A	Aggregating Multiple Models	A single model w/ Self-Following
Intermediate Supervision	N/A	N/A	N/A	Distant Supervision

# Evaluation Setup

## OpenCSR Dataset

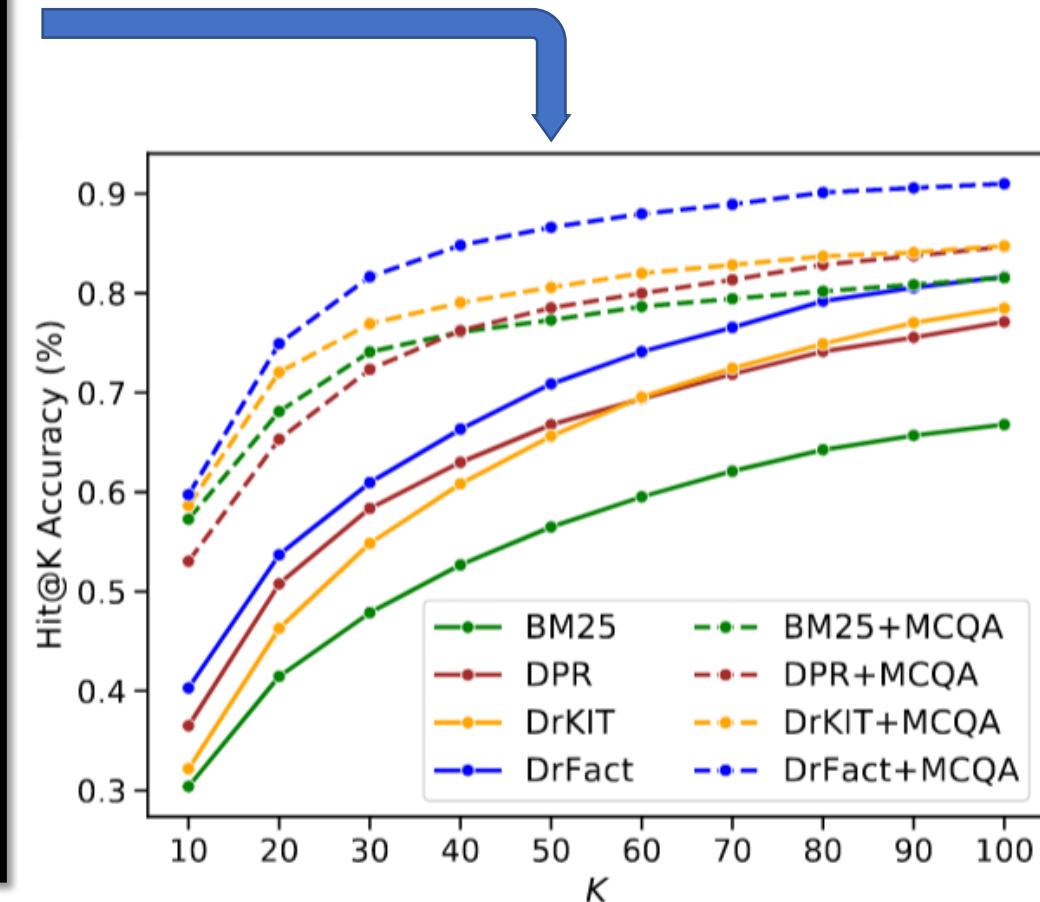
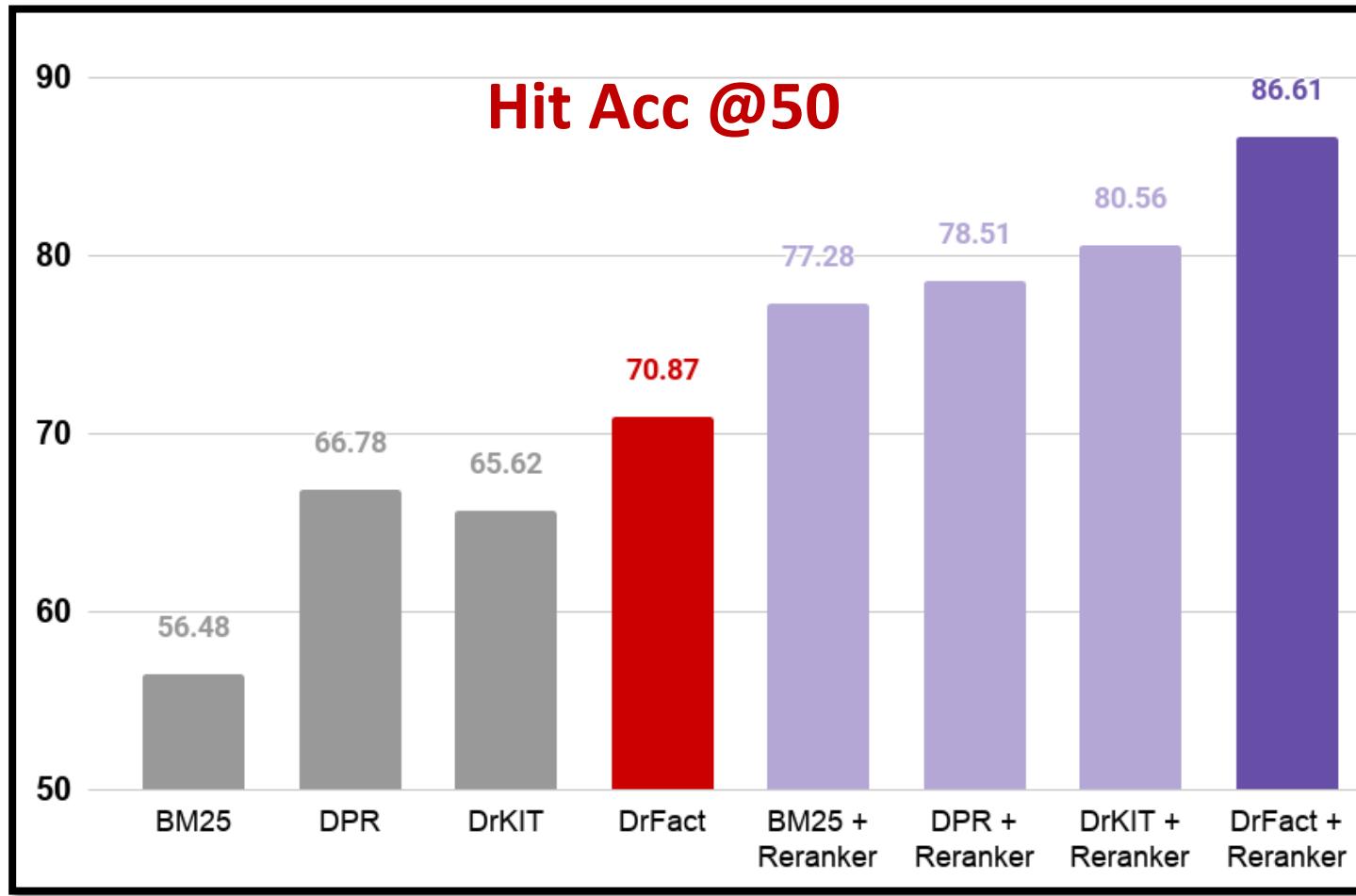
- Reformatted 3 Multiple-Choice QA Datasets (ARC, OBQA, QASC)
- + Human-Annotated Answers (7 answer concepts per question on average)



## Metrics for evaluation

- Hit Acc @ K <-- In the top K retrieved facts, there is at least one fact containing a correct answer (1 or 0).
- Ret Acc @ K <-- In the top K retrieved facts, the percentage of the recalled answer concepts (over all the correct answer concepts).
- Both are reported as an average over all examples in the test set.

# Main Experimental Results



# Experimental Results

Methods	Major Computations			Speed (sec/q)
BM25	Sparse Retrieval			0.14
DPR	BERT-base + MIPS			0.08
DrKIT	BERT-base + $T^*(\text{MIPS} + \text{sp}_{e2m})$			0.47
DRFACT	BERT-base + $T^*(\text{MIPS} + \text{sp}_{f2f})$			0.23
X + MCQA	X + $K * \text{BERT-Large}$			+ 14.12

	ARC	QASC	OBQA	Overall
$T=1$	69.3%	70.1%	65.0%	68.1%
$T=2$	71.1%	72.2%	68.3%	70.5%
$T=3 \checkmark$	71.6%	72.0%	69.0%	70.9%
w/o. Self-follow	70.9%	70.4%	68.4%	69.9%
w/o. Aux. loss	70.6%	70.1%	68.0%	69.6%

## DPR vs DrFact: Faithfulness and Interpretability

Q: "What will separate iron filings from sand?"

f1= angle irons reinforce the thinnest section of the ring ."

f2= sieves are used for separating fossils from sand..."

f3= stainless steel has a rough surface just after filing ." **DPR**

**DrFact**  
iron filings show the *magnetic fields* . (in F0)

**DrFact**  
*magnets* produce a magnetic field with a north ... (in F1)

**DrFact**  
*magnets* attract magnetic metals through magnetism (in F2)

# Findings and Take-Home Messages

- **OpenCSR is a novel setting to study CSR**
  - more **realistic** and **challenging** .
  - new **data** annotation for evaluation.
- **DrFact is an effective and efficient method for OpenCSR.**
  - **differentiable** Fact-Follow operation for **end-to-end** learning.
  - **state-of-the-art** performance comparing to strong baselines.
  - improve the explanations for multi-hop questions.

# Outline

- ~~Background of Commonsense Reasoning in NLP~~
- ~~Incorporating Structured Commonsense Knowledge~~
- ~~Incorporating Unstructured Commonsense Knowledge~~
- → **Incorporating Commonsense Knowledge for NLG**
  - **CommonGen (EMNLP 2020)**
    - A constrained text generation benchmark requires CS knowledge
    - Recent knowledge-augmentation methods for CommonGen
      - **KG-BART (AAAI 2021), KGR^4 (AAAI2022), I&V (ICLR 2022), etc.**

# 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. [Humans]
- 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:

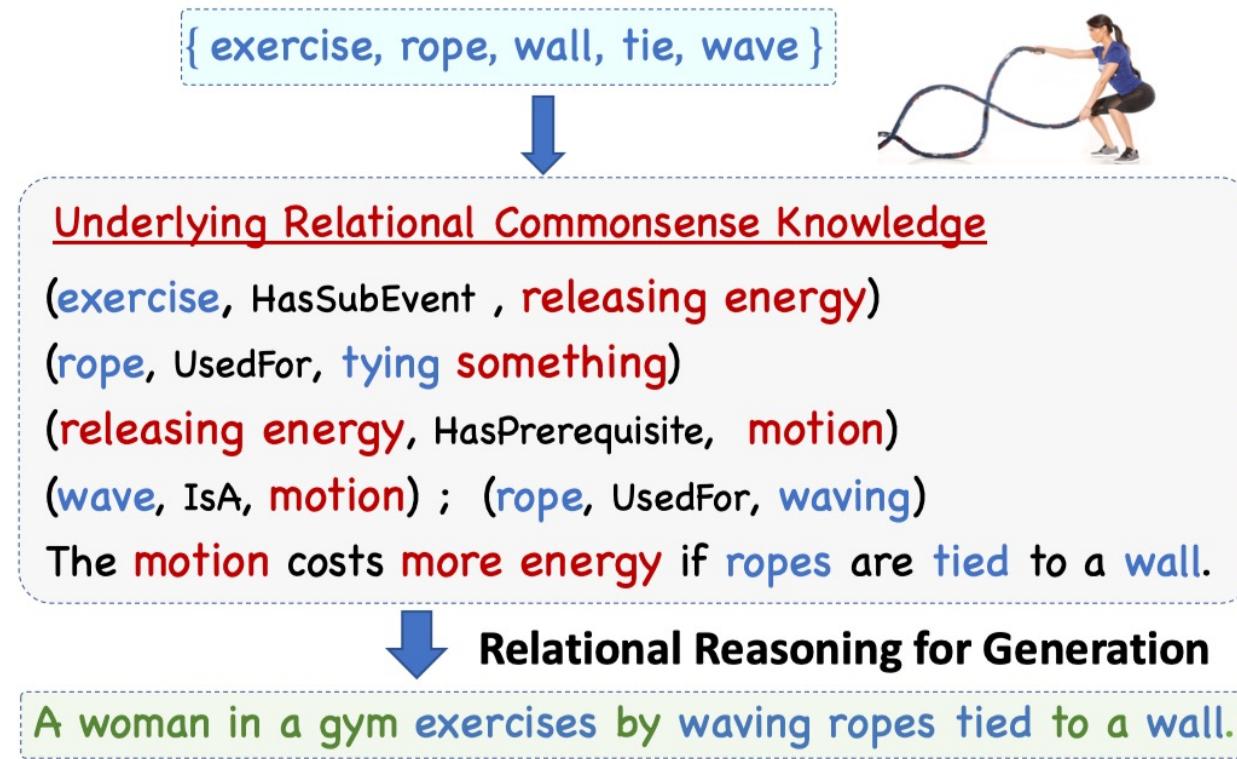
-A set of common concepts (actions & objects)

## 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**.



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.

$x_1 = \{ \text{apple}, \text{bag}, \text{put} \}$	Training
$y_1 = \text{a girl puts an apple in her bag}$	
$x_2 = \{ \text{apple}, \text{tree}, \text{pick} \}$	
$y_2 = \text{a man picks some apples from a tree}$	
$x_3 = \{ \text{apple}, \text{basket}, \text{wash} \}$	
$y_3 = \text{a boy takes an apple from a basket and washes it.}$	
	<b>Compositional Generalization</b>
$x = \{ \text{pear}, \text{basket}, \text{pick}, \text{put}, \text{tree} \}, \quad y = ?$	
<b>Reference:</b> "a girl picks some pear from a tree and put them in her basket."	Test

→ Unseen Concept in Training

Statistics	Train	Dev	Test
<b># Concept-Sets</b>	<b>32,651</b>	<b>993</b>	<b>1,497</b>
-Size = 3	25,020	493	-
-Size = 4	4,240	250	747
-Size = 5	3,391	250	750
<b># Sentences per Concept-Set</b>	67,389	4,018	6,042
Average Length	2.06	4.04	4.04
	10.54	11.55	13.34
<b># Unique Concepts</b>	4,697	766	1,248
<b># Unique Concept-Pairs</b>	59,125	3,926	8,777
<b># Unique Concept-Triples</b>	50,713	3,766	9,920
<b>% Unseen Concepts</b>	-	6.53%	8.97%
<b>% Unseen Concept-Pairs</b>	-	96.31%	100.00%
<b>% Unseen Concept-Triples</b>	-	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
	8.78	28.08	11.90	7.10	15.50	4.61	14.60	49.06
	9.66	31.14	10.70	6.10	16.40	5.06	17.20	55.70
	10.58	32.23	19.70	11.60	20.10	7.54	19.00	63.81
	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	<b>43.87</b>	<u>38.30</u>	<u>27.70</u>	29.70	<u>14.85</u>	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)	<b>22.23</b>	41.98	36.30	26.30	<b>30.90</b>	13.92	<u>30.60</u>	<b>97.35</b>
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)	<u>22.01</u>	<u>42.97</u>	<b>39.00</b>	<b>28.60</b>	<u>30.10</u>	<b>14.96</b>	<b>31.60</b>	<u>95.29</u>
Human Performance	48.88	63.79	48.20	44.90	36.20	43.53	63.50	99.31

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

(1)  
Seq2seq  
models

(2)  
Fine-tuning  
pre-trained  
LMs

(3)  
Agreement

# 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.

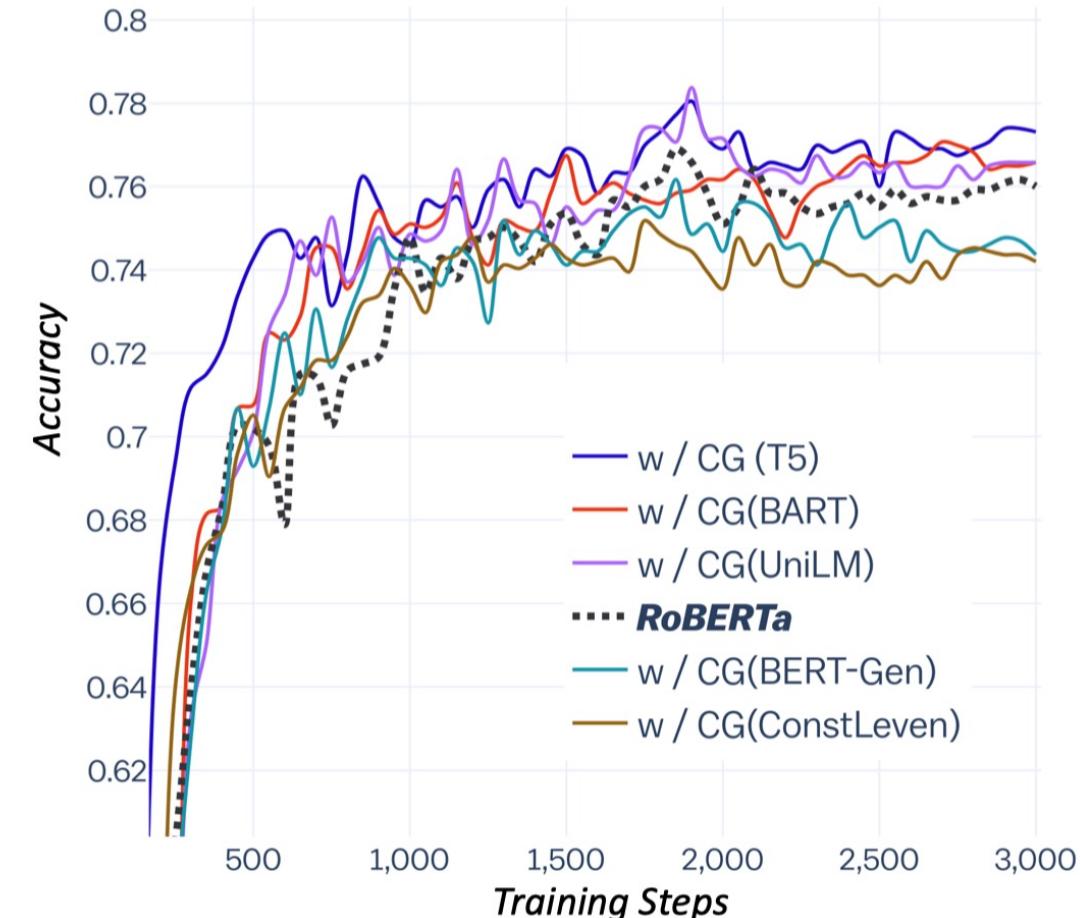
[UniLM]: hands washing soap in the sink

[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.

Rank	Model	BLEU-4	CIDEr	<b>SPICE</b>
	<b>Upper Bound</b>	<u>46.49</u>	<u>37.64</u>	<u>52.43</u>
1  Jun 09, 2021	<b>KFCNet</b> <i>MSRA and Microsoft Ads</i> <a href="#">Email</a> <a href="#">Paper (EMNLP'21)</a>	43.619	18.845	33.911
2  May 18, 2021	<b>KGR<sup>A</sup>4</b> <i>Alibaba and Xiamen University.</i> <a href="#">Email</a> <a href="#">Paper (AAAI 2022)</a>	42.818	18.423	33.564
3  Mar 23, 2021	<b>KFC (v1)</b> <i>MSRA and Microsoft Ads</i> <a href="#">Email</a> <a href="#">Paper (EMNLP'21)</a>	42.453	18.376	33.277
4  April 25, 2021	<b>R<sup>A</sup>3-BART</b> <i>Anonymous (under review).</i> <a href="#">Email</a> <a href="#">Document (placeholder)</a>	41.954	17.706	32.961
5  July 1, 2021	<b>WittGEN + T5-large</b> <i>Anonymous (under review)</i>	38.233	18.036	31.682
6  Jan 28, 2022	<b>Imagine-and-Verbalize</b> <i>USC/ISI</i> <a href="#">Email</a> <a href="#">Paper (ICLR22)</a>	40.565	17.716	31.291

# KG-BART (AAAI 2021)

**Concept Set: {river, fish, net, catch}**

**[Expected Output]:** everyday scenarios covering all given concepts.

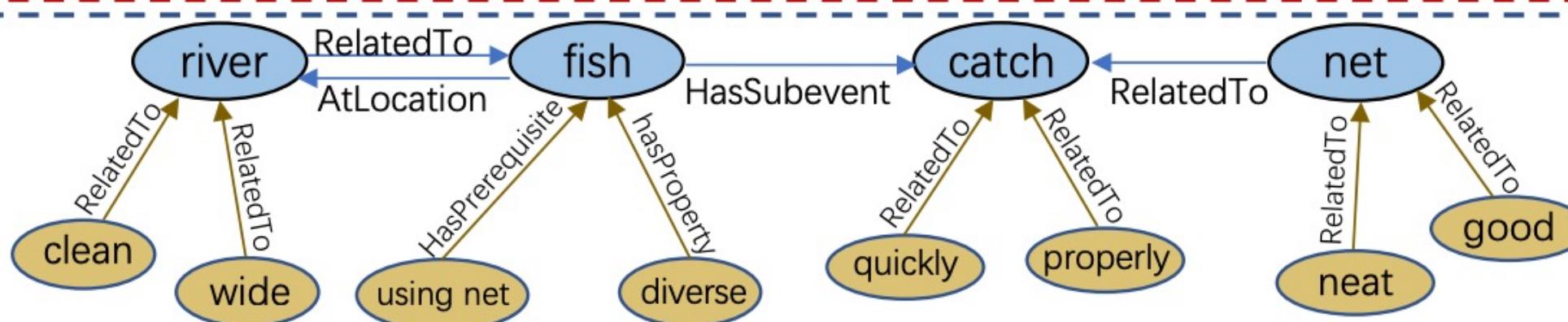
1. Fisherman uses a strong net to **catch** plentiful **fishes** in the **river**.
2. Men like to **catch fishes** in the wide **river** with a **net** in the afternoon.

**[GPT-2]:** A **fish** is catching in a **net**

**[UniLM]:** A **net** catches **fish** in a **river**

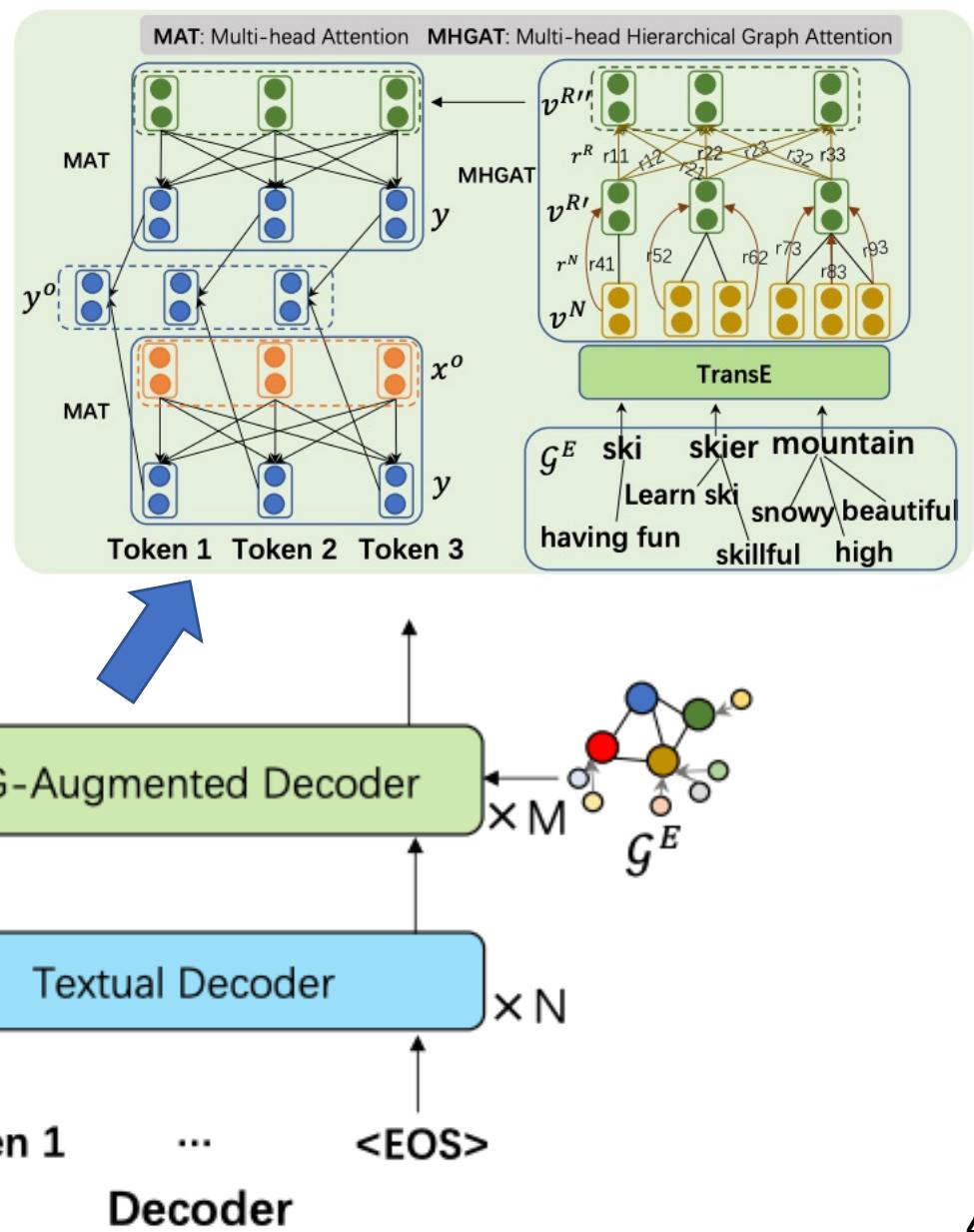
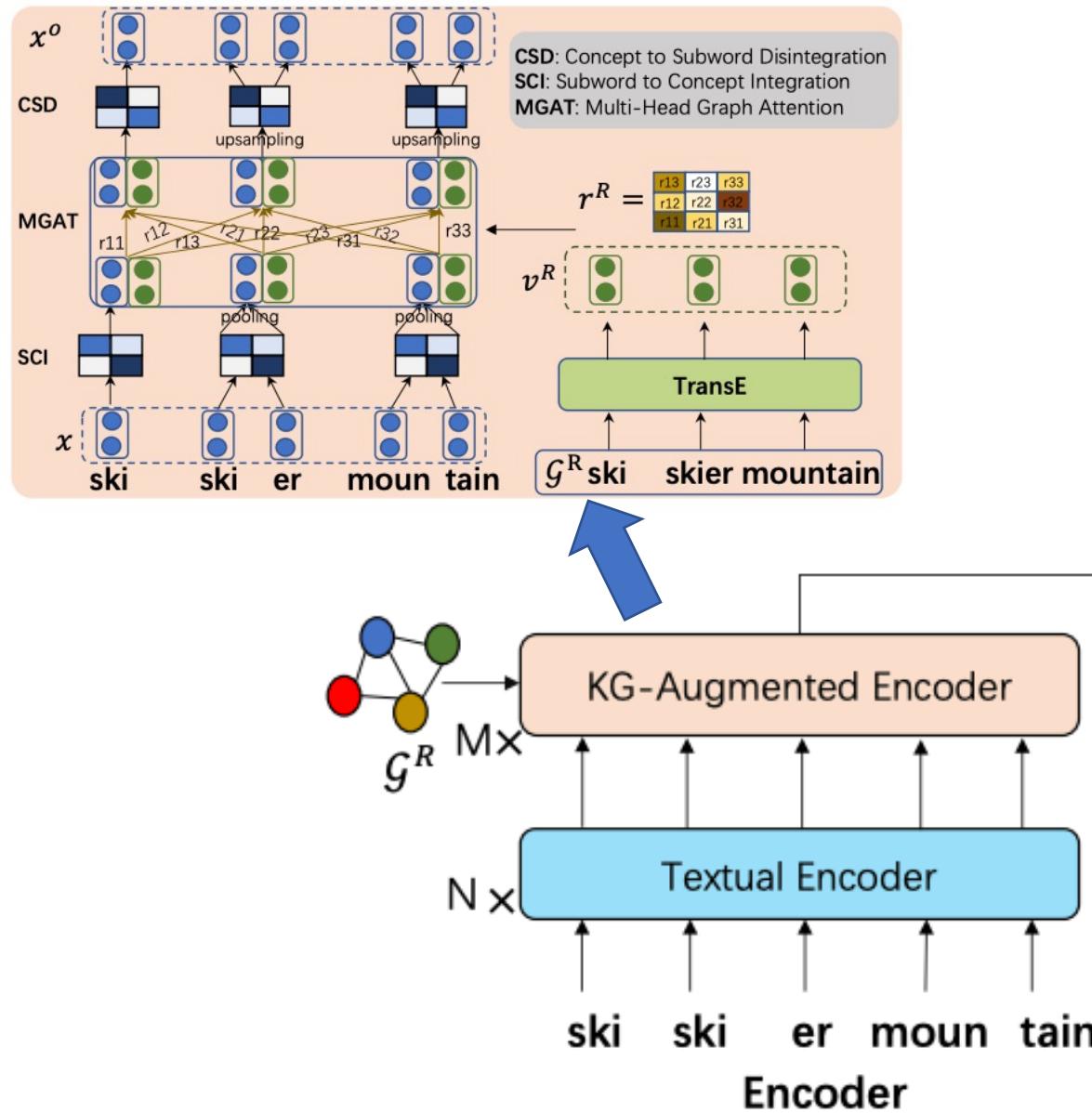
**[T5]:** **Fish** are caught in a **net** in the **river**.

**[BART]:** A man **catches a fish** with a **net** in the **river**

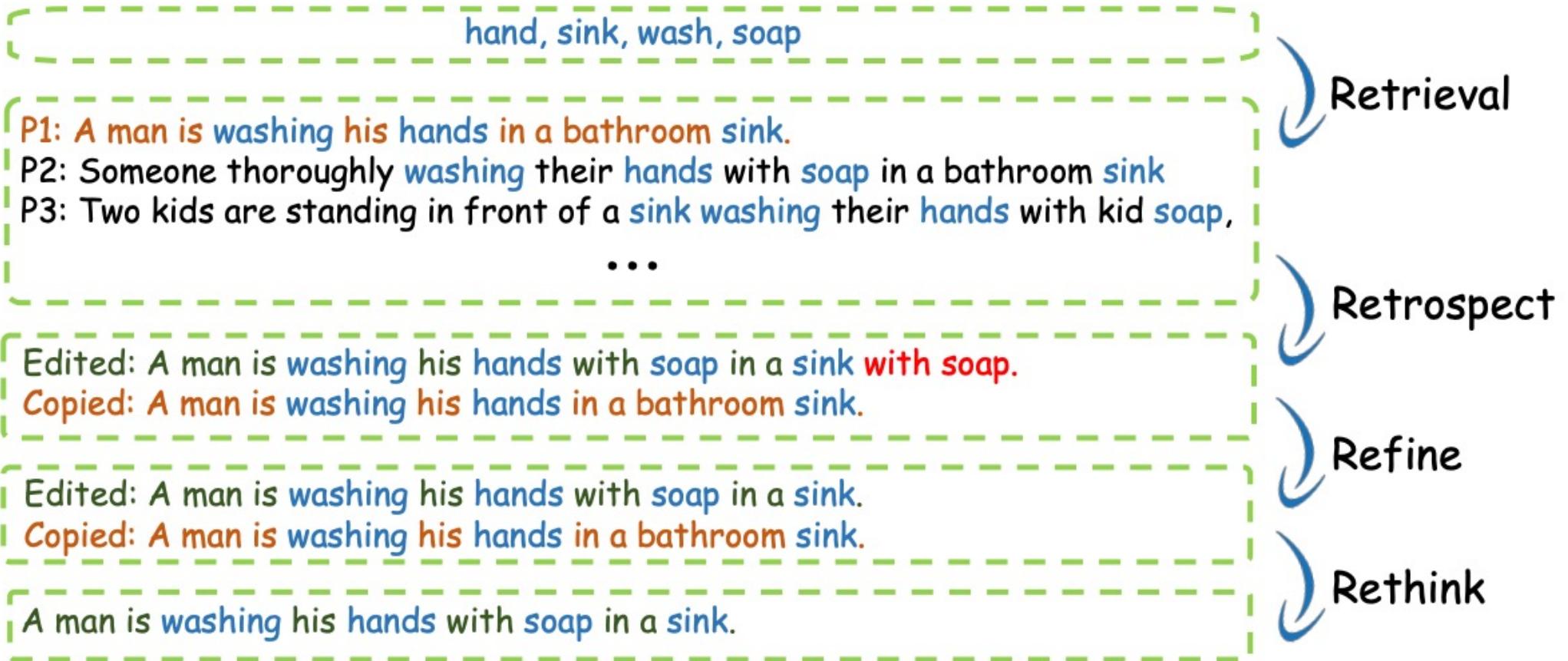
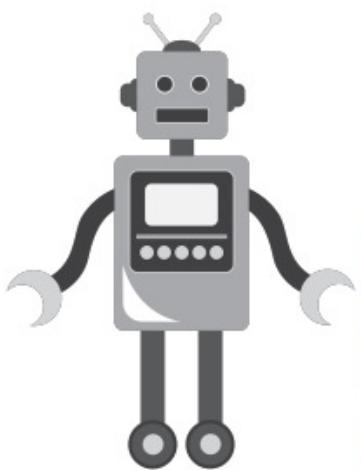


**[KG-BART]:** A fisherman **catches fishes** by using good **net** in the clean **river**.

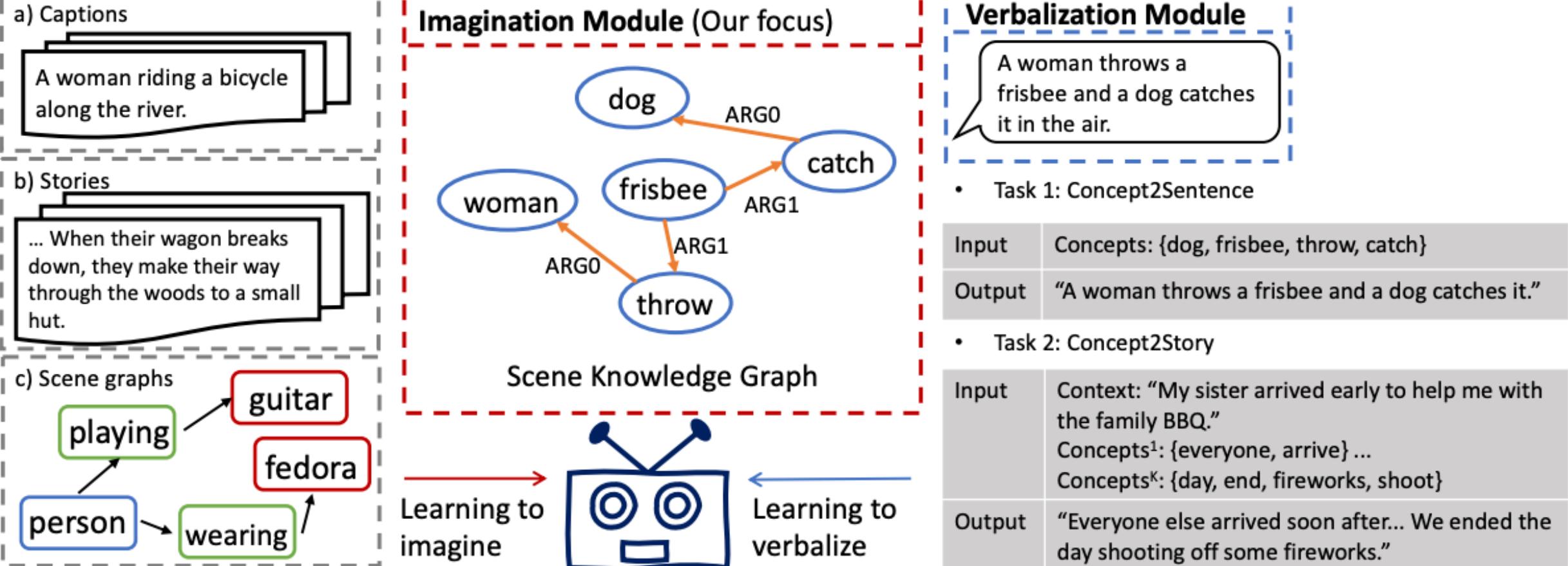
# KG-BART (AAAI 2021)



# KGR<sup>4</sup> (AAAI 2022)



# Imagine & Verbalize (ICLR 2022)



# Review

- Background of Commonsense Reasoning in NLP
- Incorporating Structured Commonsense Knowledge
  - KagNet (EMNLP 2019), MHGRN (EMNLP 2020), QA-GNN (NAACL 2021)
  - ~~GreaseLM (ICLR 2022), GSC (ICLR 2022)~~
- Incorporating Unstructured Commonsense Knowledge
  - DrFact (NAACL 2021)
- Incorporating Commonsense Knowledge for NLG
  - CommonGen (EMNLP 2020)
  - KG-BART (AAAI 2021), KGR<sup>4</sup> (AAAI2022), I&V (ICLR 2022), etc.