



---

ACL 2022 Tutorial

# Knowledge-Augmented Methods for Natural Language Generation

Meng Jiang and Wenhao Yu  
University of Notre Dame

# Related Materials



## Survey:

### A survey of knowledge-enhanced text generation

[W Yu](#), [C Zhu](#), [Z Li](#), [Z Hu](#), [Q Wang](#), [H Ji](#)... ACM Computing Surveys (CSUR), 2022 - dl.acm.org

The goal of text generation is to make machines express in human language. It is one of most important yet challenging tasks in natural language processing (NLP). Since 2014,

[☆ Save](#) [✉ Cite](#) [Cited by 38](#) [Related articles](#) [All 6 versions](#) [»](#)

- A survey of knowledge-enhanced text generation. In ACM Computing Survey
- DOI: <https://dl.acm.org/doi/10.1145/3512467>

### Knowledge-enriched Text Generation Survey, Tutorial and Reading

## Reading List:



This repository contains a list of tutorials, papers, codes, datasets, leaderboards on the topic of **Knowledge-enhanced text generation**. If you found any error, please don't hesitate to open an issue or pull request.

- Github: <https://github.com/wyu97/KENLG-Reading>

# Text generation is everywhere in our life!



- Leveraging machine intelligence to make many things EASY and FAST



*What should I wear outside today?*

*The temperature will be 85 at noon.  
You can wear short sleeve shorts.*



## Dialog systems

*“... (Two years after, Maria was born ... ) ...  
Her brother gave her a hug. ...”*



*“... 她的哥哥给了她一个大大的拥抱. ...”*

## Machine Translation

*“... Her husband was one of  
17 people killed in January’s  
terror attacks in Paris. ...  
Valerie Braham said to the  
assembled crowd. ...”*



*“... Philippe  
Braham was  
killed in the  
January’s terror  
attacks. ...”*

## Summarization

*“... The game ended with the umpire  
making a bad call, and if the call had  
gone the other way, the Blue Whales  
might have actually won the game. It  
wasn’t a victory, but I say the Blue  
Whales look like they have a shot at  
the championship, especially if they  
continue to improve.”*



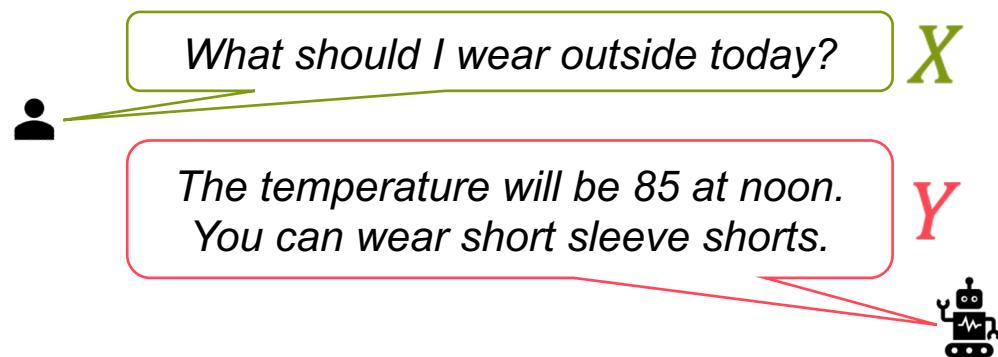
*“... The match ended with the referee  
calling wrongly, and if the call went  
the other way, the Blue Whales could  
already win the match. It was not a  
victory, but I say that the Blue Whales  
seem to have a chance in the  
championship, especially if they keep  
improving.”*

## Paraphrasing

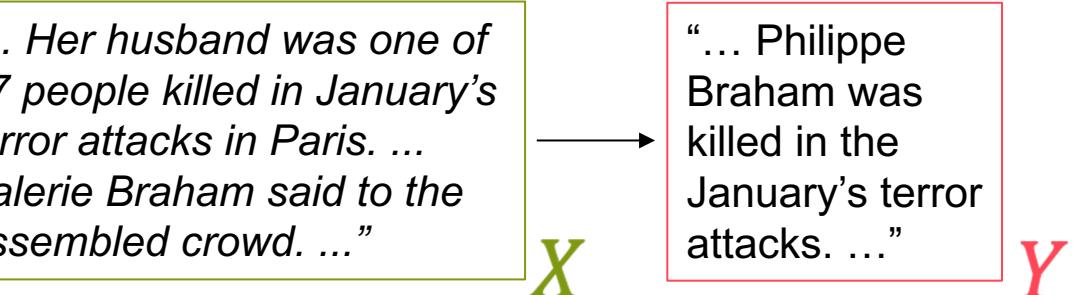
# Text generation is everywhere in our life!



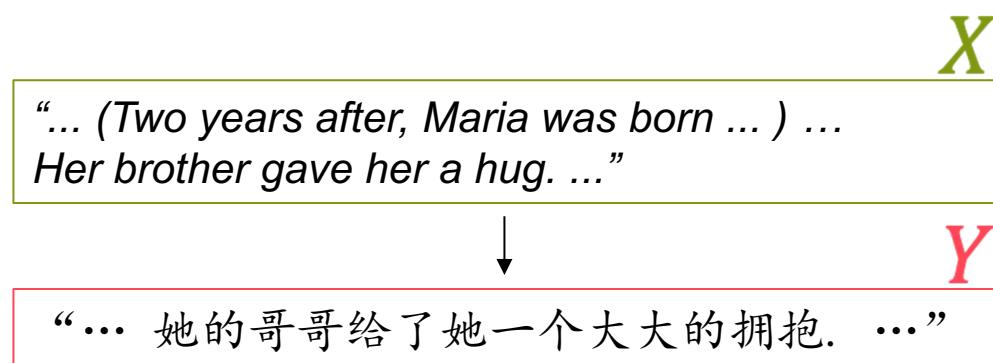
- $P(Y|X) = P(y_1, \dots, y_m|x_1, \dots, x_n) = \prod_{t=1}^m p(y_t|X, y_1, \dots, y_{t-1})$ , when **Y** is text and **X** is text.



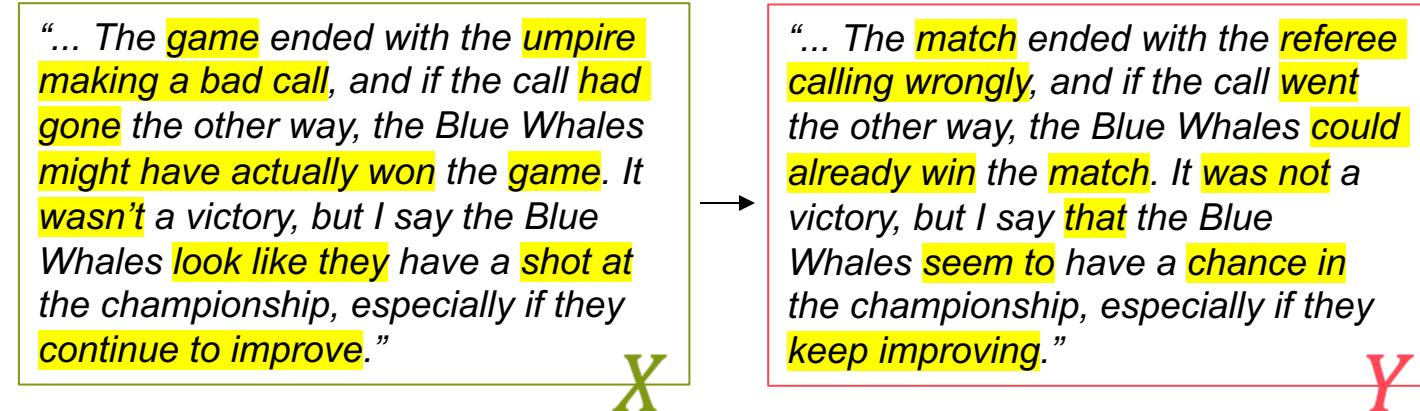
## Dialog systems



## Summarization

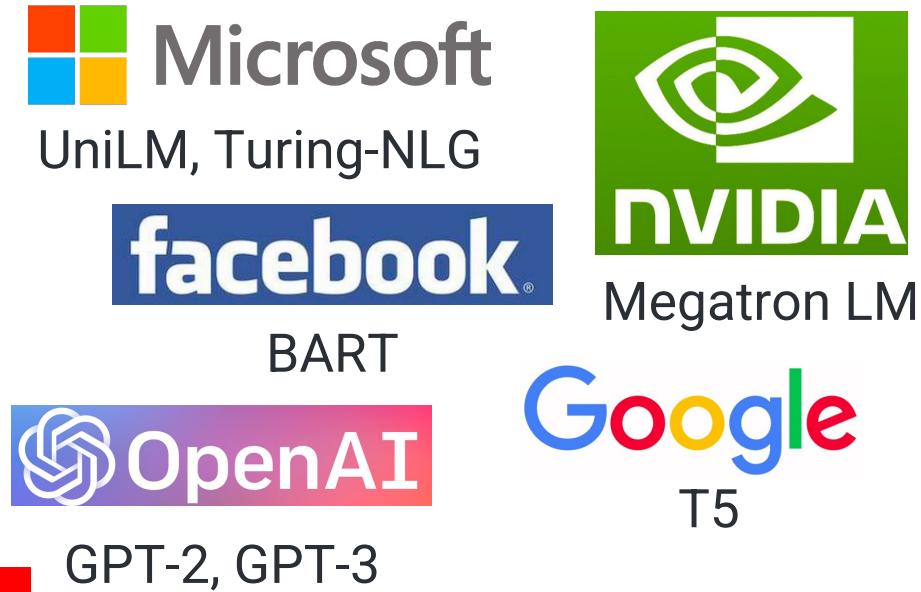
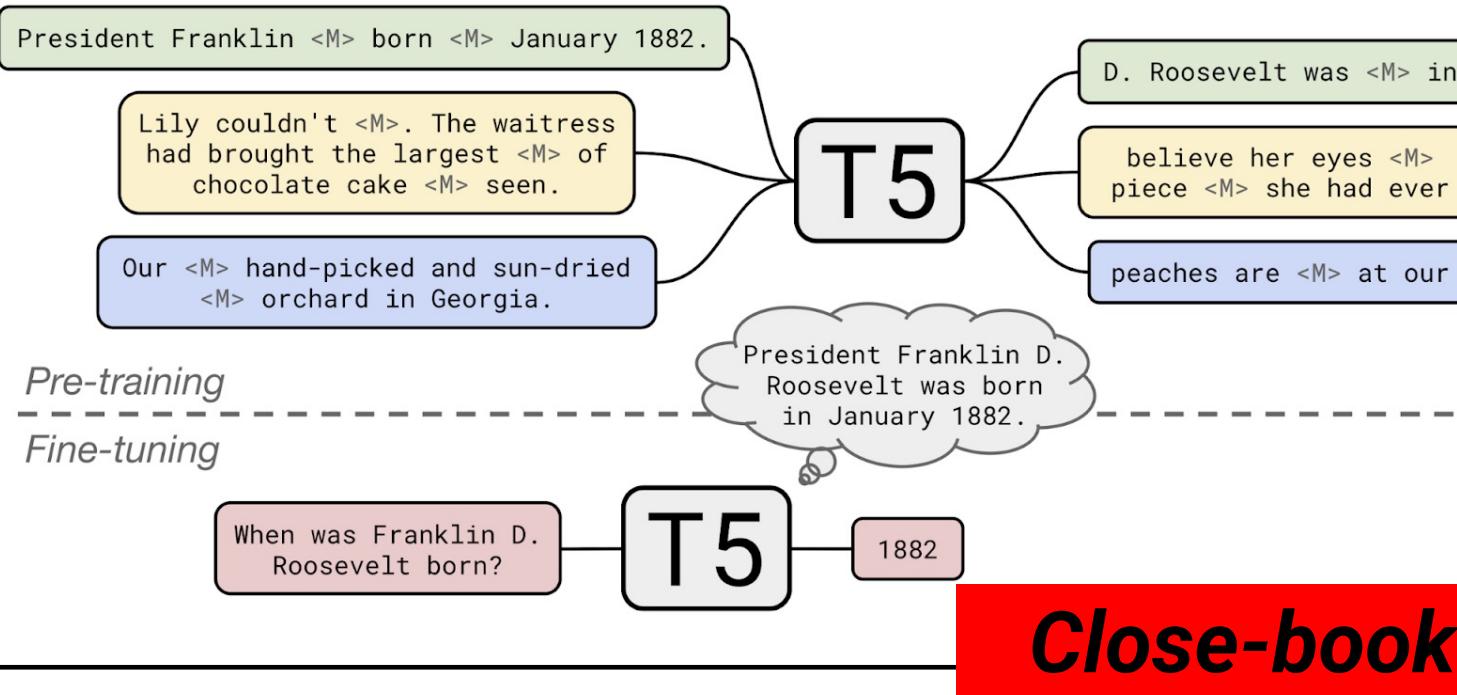


## Machine Translation

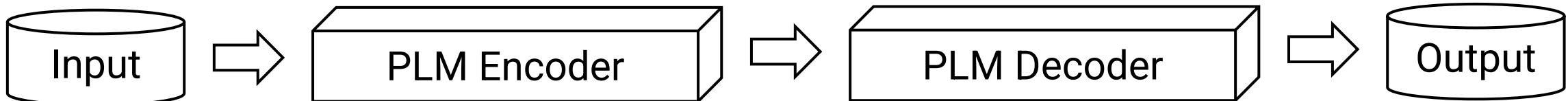


## Paraphrasing

# Fine-tune PLMs on Downstream Tasks



- Fine-tuning PLMs with *input-output* pairs of target data is the dominant paradigm in NLP research.



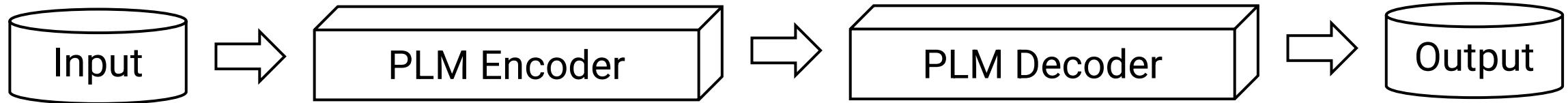
Input query: Miami Beach in Florida borders which ocean?

Output Answer: Atlantic Ocean

# Fine-tune PLMs on Downstream Tasks



-- Fine-tuning PLMs with *input-output* of target data is the dominant paradigm.



- **Machine Translation**

Input: Miami Beach in Florida borders  
which ocean?

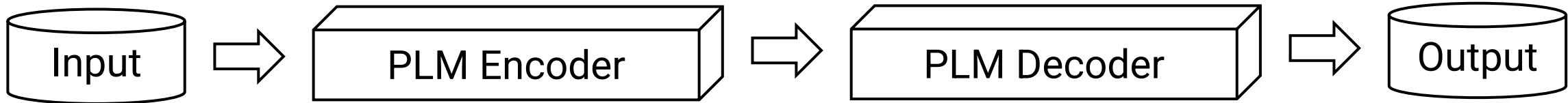
Output: 佛罗里达州的迈阿密海滩与  
哪个海洋接壤? (from Google Translate)

The screenshot shows the Google Translate mobile application interface. At the top, it says "Google Translate" with a menu icon and a user profile picture. Below that are tabs for "Text" and "Websites", and language selection buttons for "ENGLISH" and "CHINESE (SIMPLIFIED)". The input text in English is "Miami Beach in Florida borders which ocean?". The output text in Chinese is "佛罗里达州的迈阿密海滩与哪个海洋接壤?". Below the text, there is a phonetic transcription in Pinyin: "Fóluólídá zhōu de mài'āmì hǎitān yǔ nǎge hǎiyáng jiērāng?". The bottom of the screen shows navigation icons for back, forward, and search.

# Fine-tune PLMs on Downstream Tasks



-- Fine-tuning PLMs with *input-output* of target data is the dominant paradigm.



- **Summarization**

**Input:** BERT paper full text

**Output:** TLDR (from semantic scholar)

(A new language representation model, BERT, designed to pre-train deep bidirectional representations from unlabeled text by jointly conditioning both left and right context in all layers)

## BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding

Jacob Devlin, Ming-Wei Chang, Kenton Lee, Kristina Toutanova • Computer Science • NAACL • 2019

**TLDR** A new language representation model, BERT, designed to pre-train deep bidirectional representations from unlabeled text by jointly conditioning on both left and right context in all layers, which can be fine-tuned with just one additional output layer to create state-of-the-art models for a wide range of tasks.

### Abstract

We introduce a new language representation model called **BERT**, which stands for Bidirectional Encoder Representations from Transformers. Unlike recent language representation models (Peters et al., 2018a; Radford et al., 2018), BERT is designed to pre-train deep bidirectional representations from unlabeled text by jointly conditioning on both left and right context in all layers. As a result, the pre-trained BERT model can be fine-tuned with just one additional output layer to create state-of-the-art models for a wide range of tasks, such as question answering and language inference, without substantial task-specific architecture modifications. BERT is conceptually simple and empirically powerful. It obtains new state-of-the-art results on eleven natural language processing tasks, including pushing the GLUE score to 80.5 (7.7 point absolute improvement), MultiNLI accuracy to 86.7% (4.6% absolute improvement), SQuAD v1.1 question answering Test F1 to 93.2 (1.5 point absolute improvement) and SQuAD v2.0 Test F1 to 83.1 (5.1 point absolute improvement). [Collapse](#)

31,856

PDF

View on ACL

Save

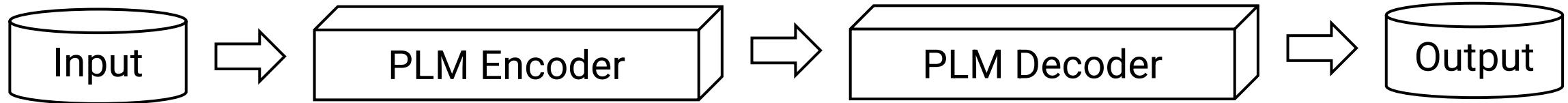
Alert

Cite

# Fine-tune PLMs on Downstream Tasks



-- Fine-tuning PLMs with *input-output* of target data is the dominant paradigm.



## • Summarization

BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding

Though the input and output are different (e.g., language, length), the contents are very similar, and globally, under the same topic.

## LITTLE SEMANTIC GAP

sentations from unlabeled text by jointly conditioning both left and right context in all layers)

MULTIEC accuracy to 86.7% (new absolute improvement) requires 111 question-answer pairs (1.6 point absolute improvement) and SQuAD v2.0 Test F1 to 83.1 (5.1 point absolute improvement). [Collapse](#)

31,856

PDF

View on ACL

Save

Alert

Cite

# Why knowledge is needed in NLG?



## Question/Answer Generation (e.g., open-domain QA, question generation)

- Query: Who did Hawaii belong to before 1895?



Question-answer gap

**Hawaiian Kingdom**



WIKIPEDIA

Hawaii is the most recent state ... **On Jan 17, 1895**. The United States Minister to the **Hawaiian Kingdom** conspired with U.S. citizens to overthrow the monarchy.

**In 1895**, United States Public Law acknowledged that “the overthrow of **Hawaiian Kingdom** occurred with the active participation of agents ...



The **Hawaiian Kingdom**, or Kingdom of Hawaii, was a sovereign state located in the Hawaiian Islands formed in 1795.



Subject: **Hawaiian Kingdom**  
Relation: end time  
Object: **1895**

# Why knowledge is needed in NLG?



## Creative Generation (e.g., story generation, abductive generation)

Storyline content semantic gap

- (ROCStories) Input:

(S1) Mr. Egg was presenting a volcanic *eruption* to the *science* class.

(S2) He has a diagram of a *volcano* that

looked like it was made of tinfoil

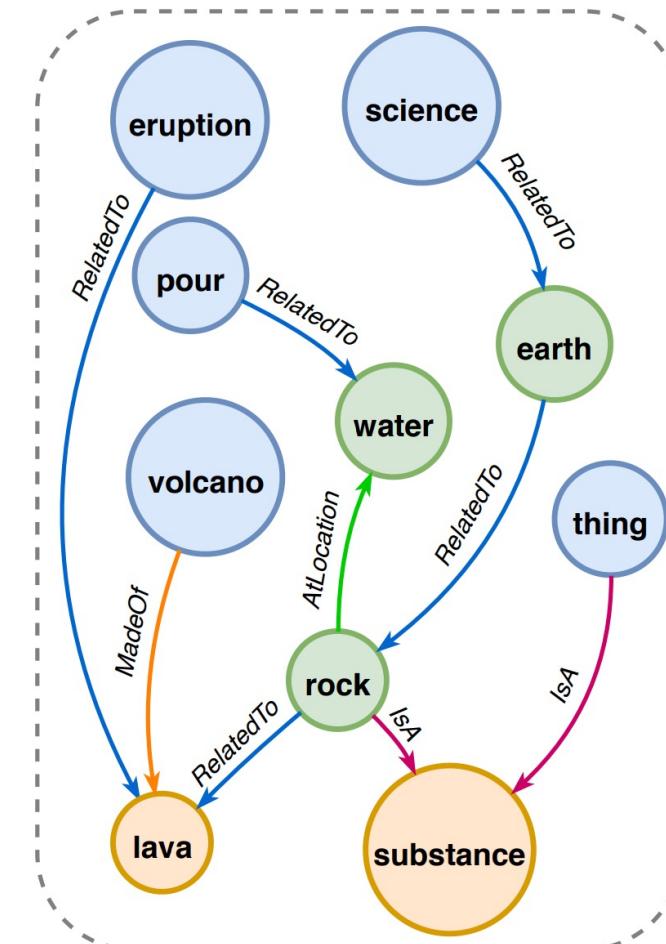
(S3) He then took out a huge *thing* of

vinegar and started to *pour* it in!

(S4) The class had no clue what was going on and looked on in astonishment.

- (ROCStories) Output:

→ (S5) The volcano then exploded with *substance* that looked like *lava*!



# Why knowledge is needed in NLG?



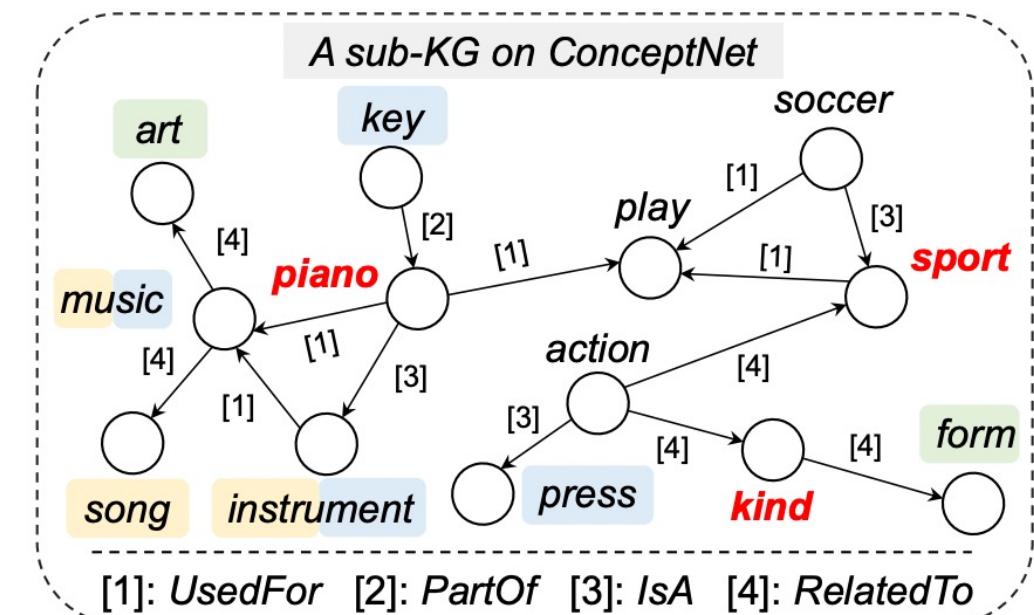
## Explanation Generation (e.g., choice explanation, counterfactual explanation)

→ **Input: Piano is a kind of sport .**

→ **Statement-explanation gap**

→ **Outputs: 3 different explanations**

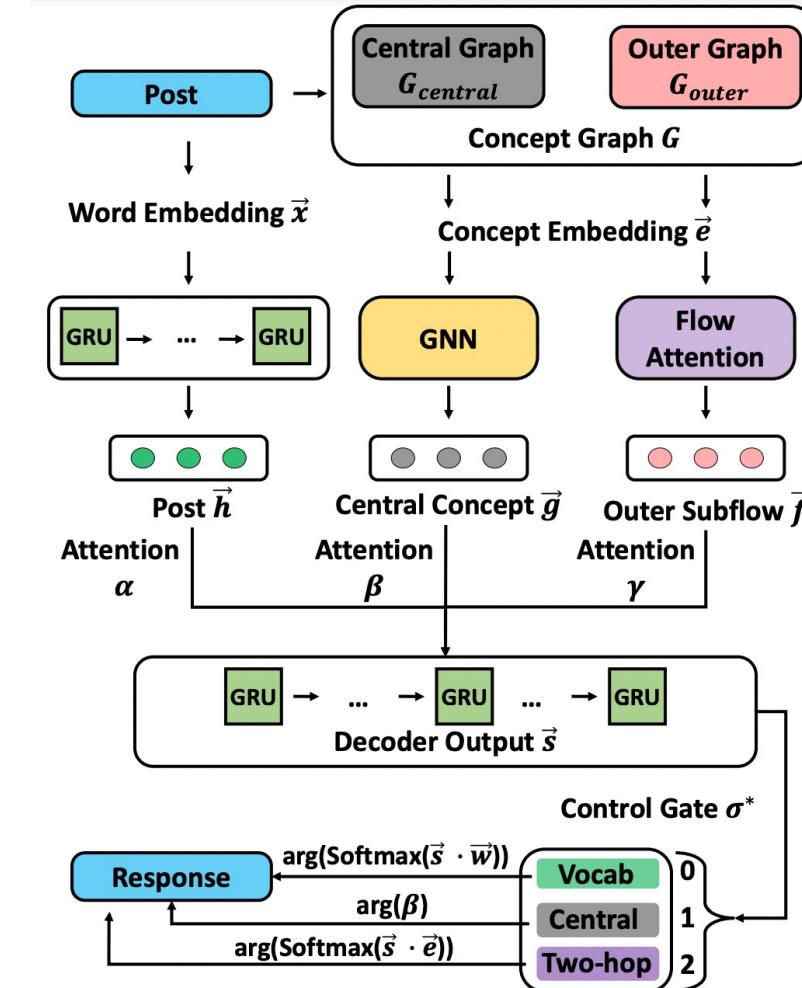
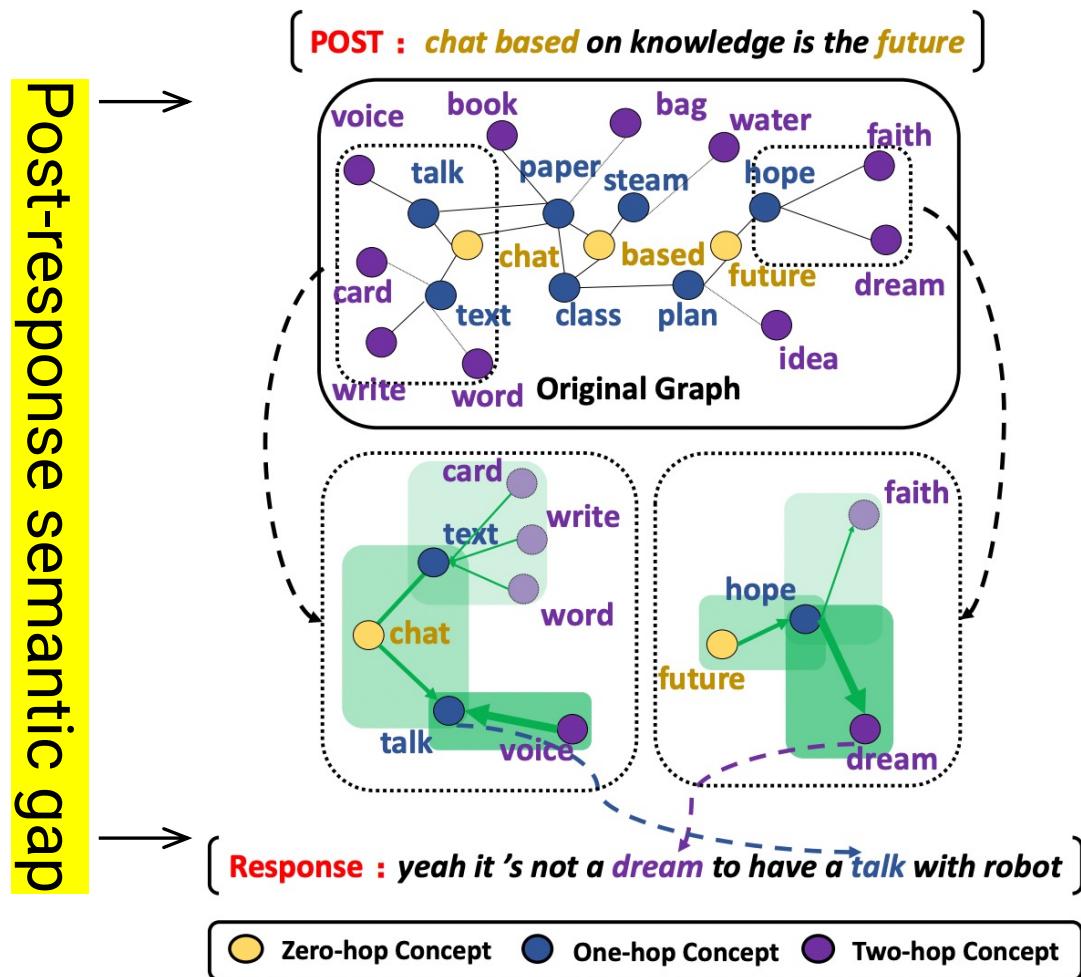
- (1) You can produce music when pressing keys on the piano, so it is an instrument .
- (2) Piano is a musical instrument used in songs to produce different musical tones .
- (3) Piano is a kind of art form .



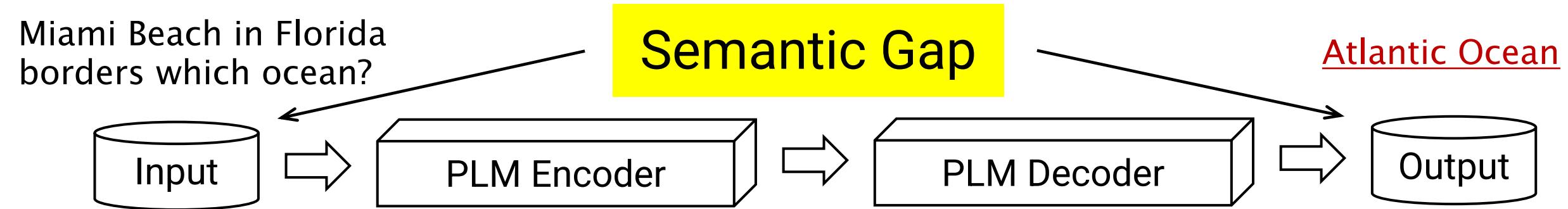
# Why knowledge is needed in NLG?



## Response Generation (e.g., chit-chat conversation, open-domain dialogue)



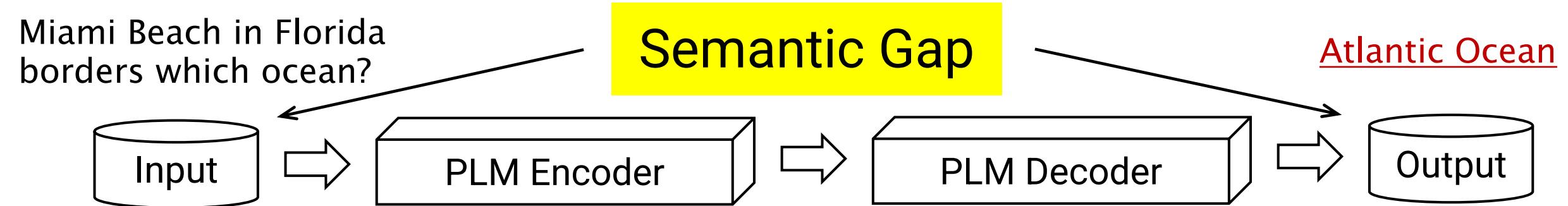
# Why Close-book does not work?



- When fine-tuned on the downstream tasks, the LMs might forget previously learned knowledge during pre-training, leading to catastrophic forgetting.

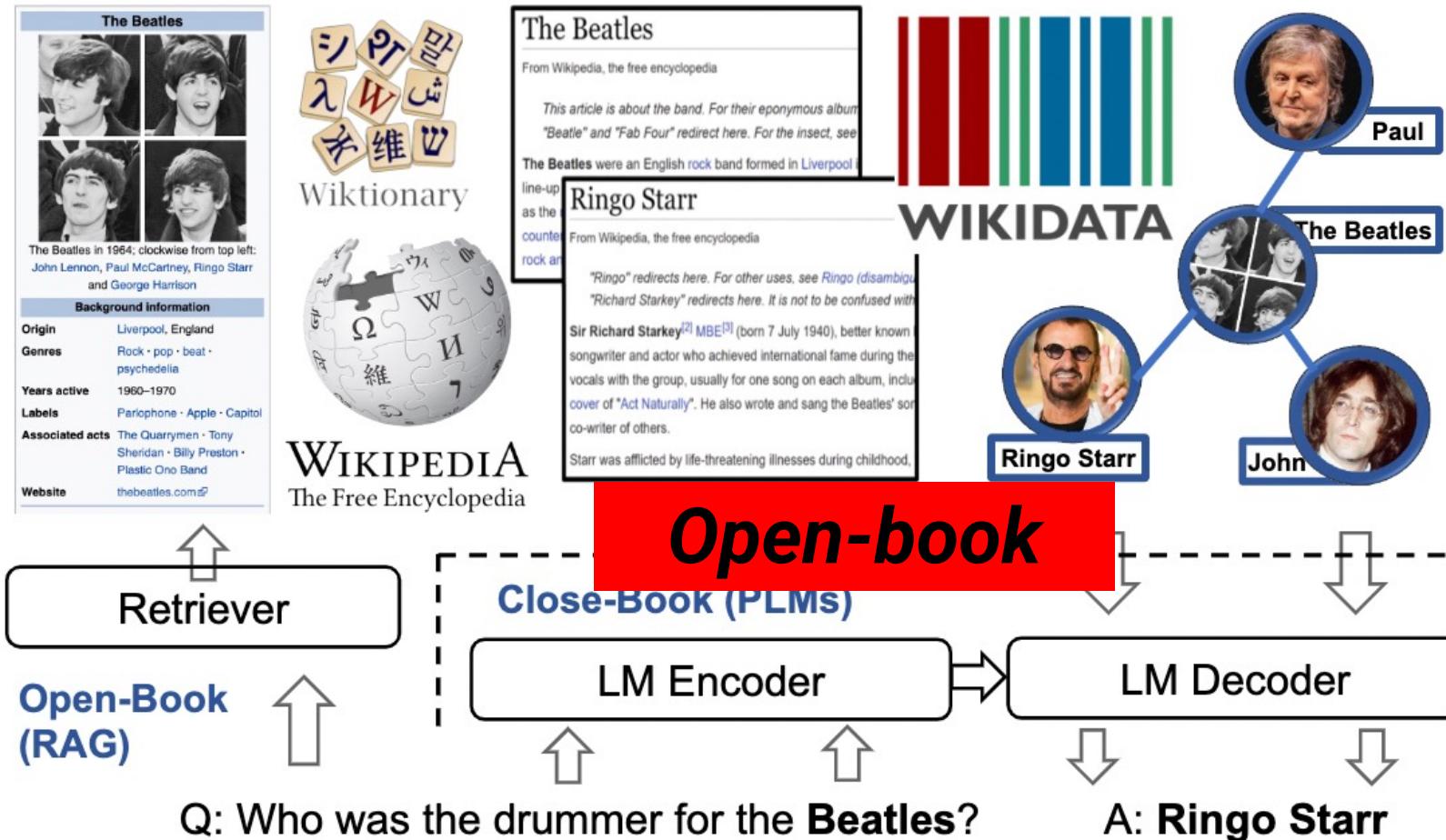
Model		Open Natural Questions			TriviaQA			WebQuestions					
		Total	Question Overlap	Answer Overlap Only	No Overlap	Total	Question Overlap	Answer Overlap Only	No Overlap	Total	Question Overlap	Answer Overlap Only	No Overlap
Closed book	T5-11B+SSM	36.6	77.2	22.2	9.4	-	-	-	-	44.7	82.1	44.5	22.0
	BART	26.5	67.6	10.2	0.8	26.7	67.3	16.3	0.8	27.4	71.5	20.7	1.6
Nearest Neighbor	Dense	26.7	69.4	7.0	0.0	28.9	81.5	11.2	0.0	26.4	78.8	17.1	0.0
	TF-IDF	22.2	56.8	4.1	0.0	23.5	68.8	5.1	0.0	19.4	63.9	8.7	0.0

# Why Close-book does not work?



- When fine-tuned on the downstream tasks, the LMs might forget previously learned knowledge during pre-training, leading to the catastrophic forgetting.
- The LMs make predictions by only “looking up information” stored in its parameters, leading to inferior performance and interpretability.
- They are usually trained offline, rendering the model agnostic to the latest information, e.g., asking BERT (released at 2018) about COVID-19.
- They are often expensive to train (e.g., GPT-3, T5-11B, GLaM-1.2T).

# Open-book: Bridge the Semantic Gap



-- **Knowledge-enhanced methods:** Knowledge is **retrieved** based on the input text. The Language model make predictions by *reading and reasoning* over retrieved information.

# Open-book: Bridge the Semantic Gap



**Structure Knowledge**  
(i.e., knowledge graph)

A Survey of Knowledge-Enhanced Text Generation  
(Yu et al., ACM Computing Survey 2022)



**Unstructured Knowledge**  
(i.e., grounded document)



**Encyclopedic Knowledge**  
(i.e., Wikipedia, AMiner)

A Survey of Knowledge-Intensive Natural Language Processing  
(Yin et al., on arXiv 2022)



**Commonsense Knowledge**  
(i.e., OMCS, ConceptNet)

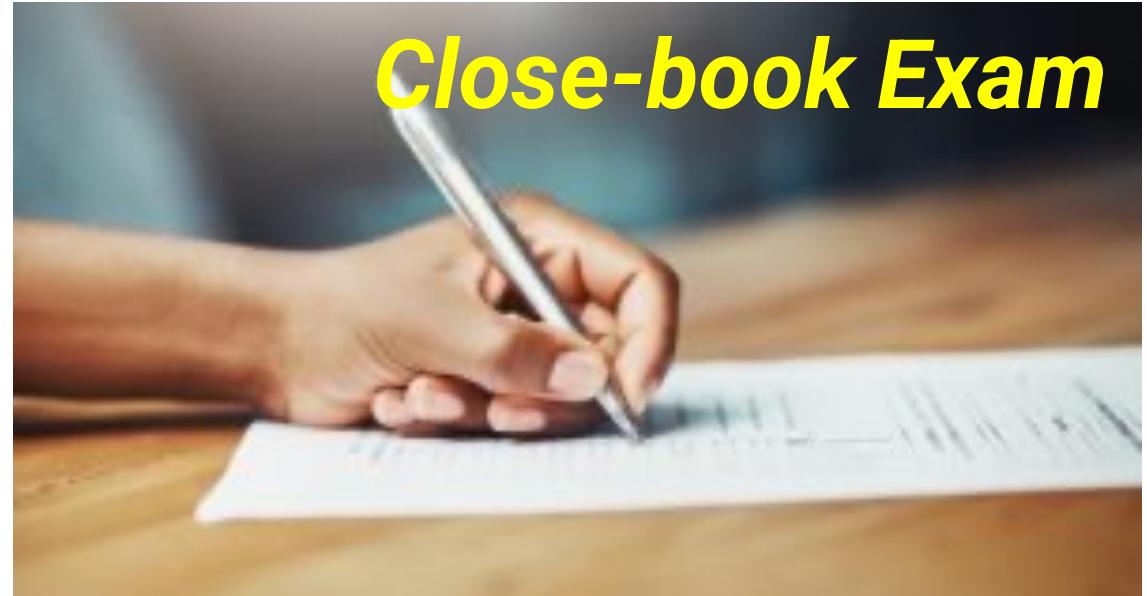
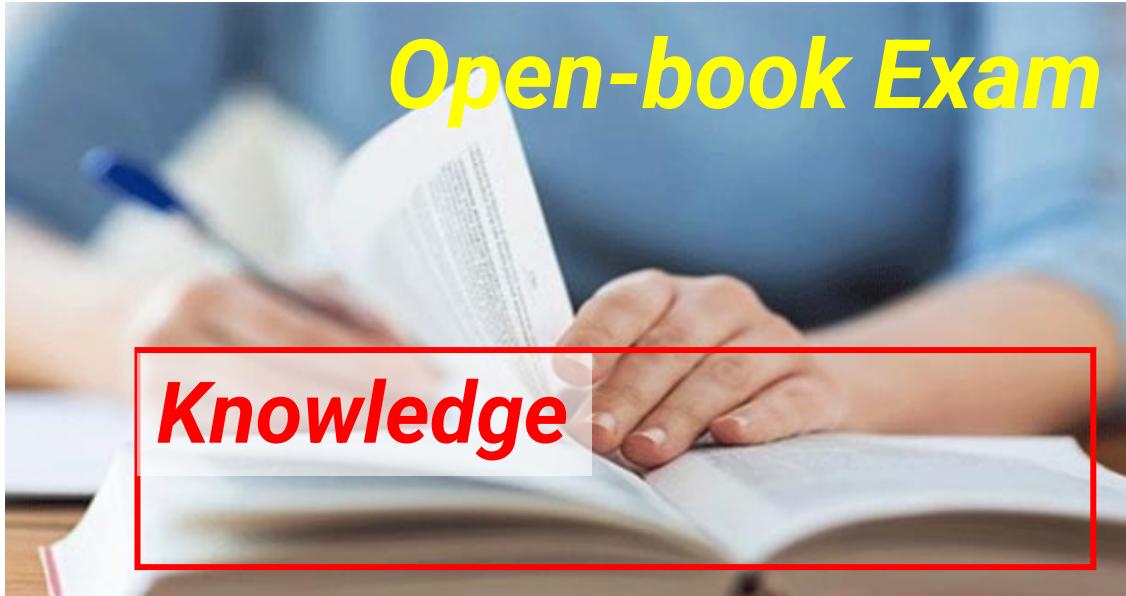
# Summary: Close-book & Open-book



**Close-book Models:** Knowledge is learnt into a language model (LM) parameters. During fine-tuning, only feed *input text* into LMs and make predictions. Examples include BART, T5, GPT-2/GPT-3, UniLM.

**Open-book Models:** We also refer it as knowledge-enhanced methods. Knowledge is retrieved based on the input text. The Language model make predictions by *reading and reasoning* over the retrieved texts. Examples include JointGT, REALM, KG-BART, RAG (*will be covered in this tutorial*).

# Open-book (pros) v.s. Close-book

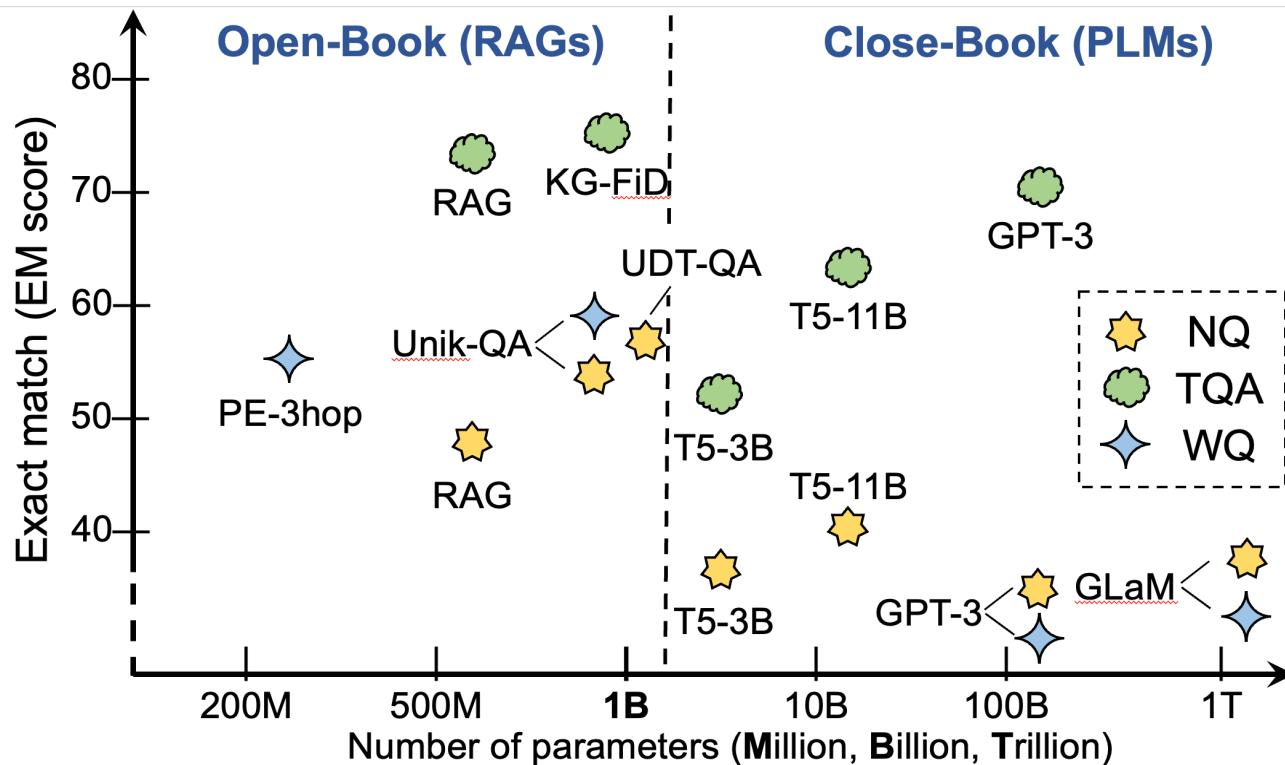


- The knowledge is not implicitly stored in model parameters, but is explicitly acquired in a plug-and-play manner, leading to great scalability and interpretability.
- Instead of generating from scratch, the paradigm generating text from some retrieved references, which potentially alleviates the difficulty of text generation.

# Open-book (SoTA) v.s. Close-book (SoTA)



- From the existing literature, the performance of the open-book method is usually significantly better than that of the close-book method. At the same time, open-book methods are trained with less parameters.



	Close-book SoTA	Open-book SoTA
Model	GLaM-1.2T	UnikQA-990M
WQ	25.3	57.8
Model	GPT-175B	KGFiD-994M
TQA	68.0	72.5
Model	T5-11B	UDTQA-990M
NQ	42.3	55.2

# Outline of Knowledge-enhanced NLG

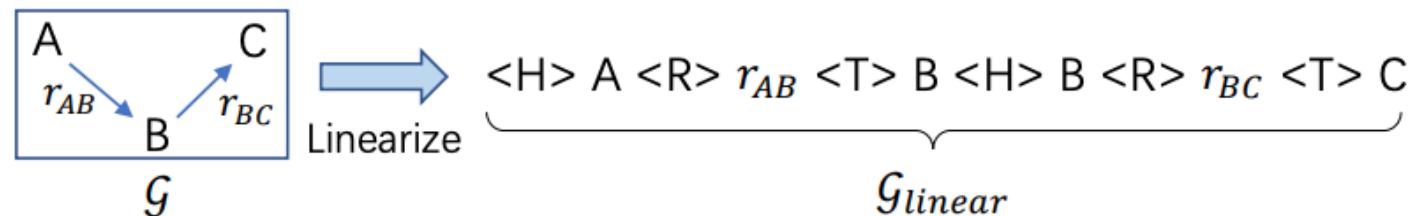


	<b>Knowledge-enhanced pre-training (self-supervised training)</b>	<b>Knowledge-enhanced fine-tuning (end-to-end training)</b>
Structured knowledge	KG-BART [AAAI 2021] JointGT [ACL 2021]	GRF [EMNLP 2020] KGMoE [ACL 2022]
Unstructured knowledge	REALM [ICML 2020]	RAG [Neurips 2020], FiD [EACL 2021], Re-T5 [ACL 2021]

- **Covered NLP applications:** Open-domain question answering, commonsense reasoning, dialogue system, question generation, data-to-text generation ...

## JointGT: Graph-Text Joint Representation Learning for Text Generation from Knowledge Graphs. ACL'21

- First work of jointly training knowledge graph and text representation for NLG
- Pre-training dataset: 7M KG-text pairs extracted from Wikipedia/Wikidata
- **Feed KG triples into Transformer:**



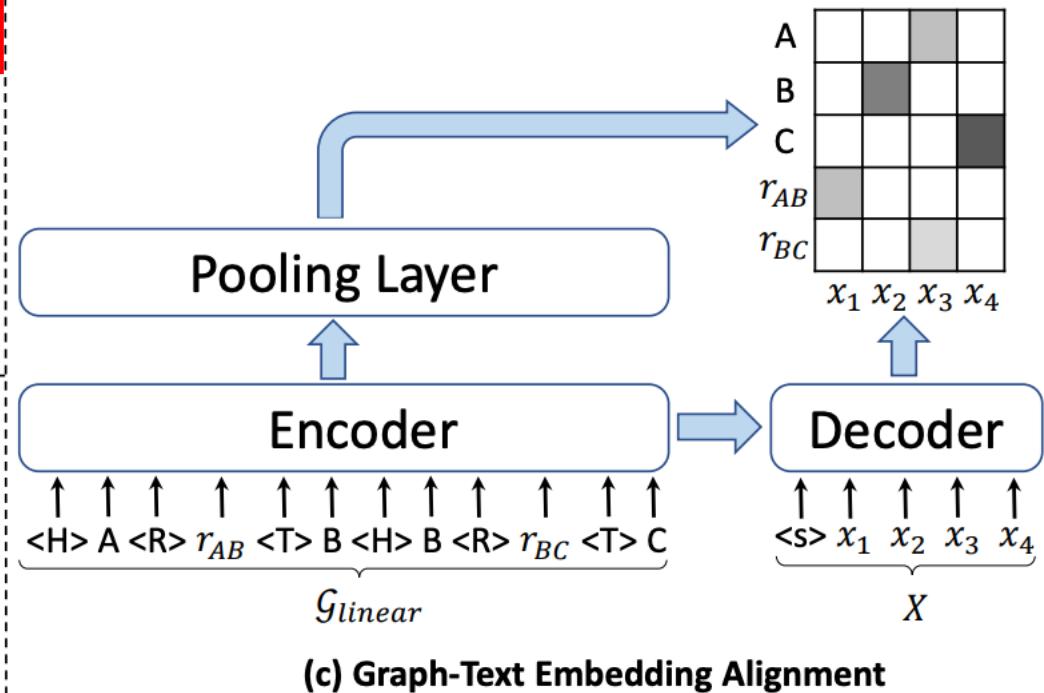
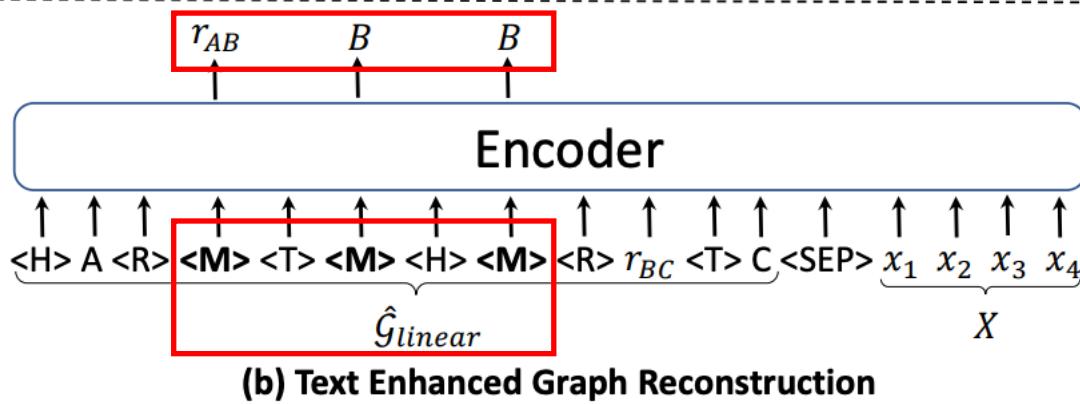
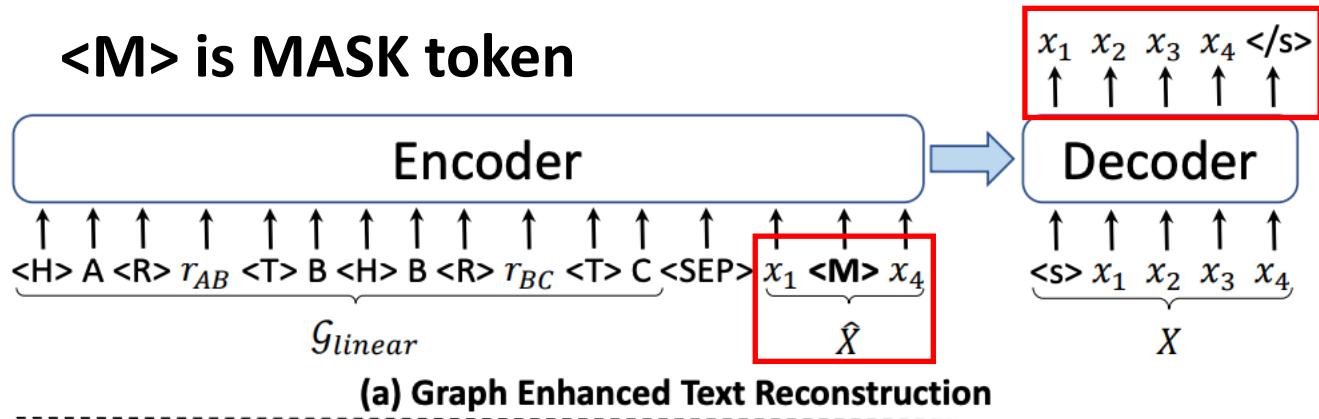
SubKG -> Sequence      <H>: head entity, <R>: relation, <T>: tail entity

- **Three novel pre-training tasks:** Graph Enhanced Text Reconstruction, Text Enhanced Graph Reconstruction, Graph-Text Embedding Alignment

## JointGT: Graph-Text Joint Representation Learning for Text Generation from Knowledge Graphs. ACL'21

- Two novel generation pre-training tasks; one graph-text alignment objective

**<M> is MASK token**



## JointGT: Graph-Text Joint Representation Learning for Text Generation from Knowledge Graphs. ACL'21

- Downstream tasks: data-to-text generation, knowledge base question generation

Dataset	#Param	WebNLG(U)			WebQuestions			PathQuestions		
		BLEU	METEOR	ROUGE	BLEU	METEOR	ROUGE	BLEU	METEOR	ROUGE
SOTA-NPT	-	61.00 <sup>†</sup>	42.00 <sup>†</sup>	71.00 <sup>†</sup>	29.45 <sup>‡</sup>	30.96 <sup>‡</sup>	55.45 <sup>‡</sup>	61.48 <sup>‡</sup>	44.57 <sup>‡</sup>	77.72 <sup>‡</sup>
KGPT	177M	64.11 <sup>#</sup>	46.30 <sup>#</sup>	74.57 <sup>#</sup>	-	-	-	-	-	-
BART	140M	64.55	46.51	75.13	29.61	31.48	55.42	63.74	47.23	77.76
T5	220M	64.42	46.58	74.77	28.78	30.55	55.12	58.95	44.72	76.58
JointGT (BART)	160M	65.92	47.15	76.10**	30.02*	32.05**	55.60	65.89**	48.25**	78.87**
JointGT (T5)	265M	66.14**	47.25**	75.91	28.95	31.29	54.47	60.45	45.38	77.59

Table: Results on three text generation tasks, including WebNLG, WebQuestions and PathQuestions.

## REALM: Retrieval-Augmented Language Model Pre-Training. ICML'20

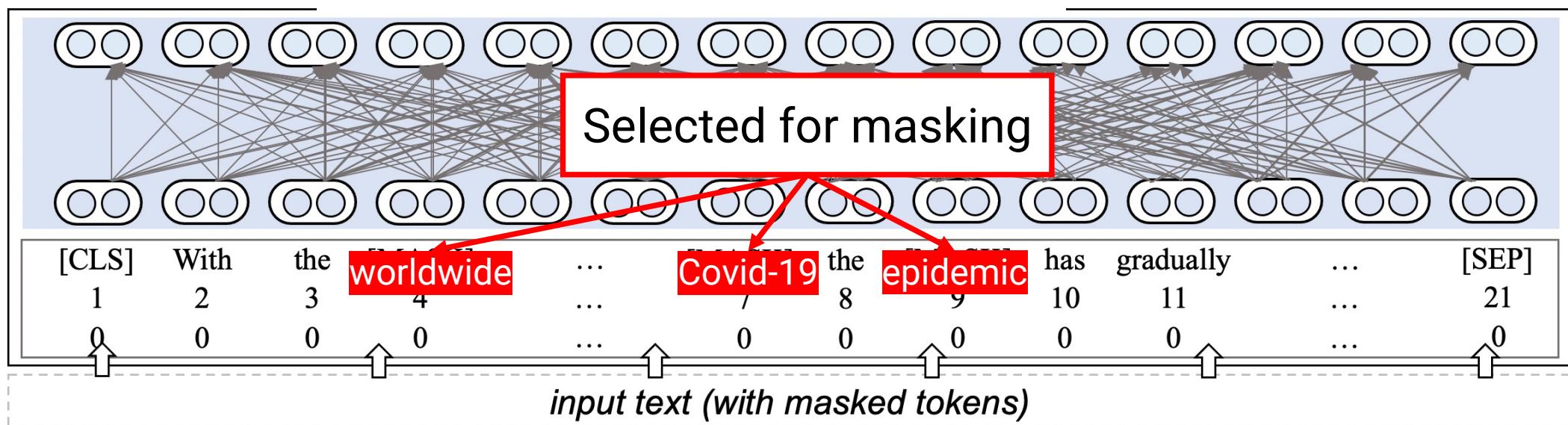
Pre-training  
tasks

BERT  
architecture

Token Emb  
Pos. Emb  
Type Emb

Input text

*Pre-training task: Masked Language Model*



**[CLS]** With the worldwide spread of **COVID-19**, the **epidemic** has gradually attracted widespread xxx. xxx. **[SEP]**

**Input text:** With the worldwide spread of COVID-19, the epidemic has gradually attracted widespread attention and discussion on social media platforms.

## REALM: Retrieval-Augmented Language Model Pre-Training. ICML'20

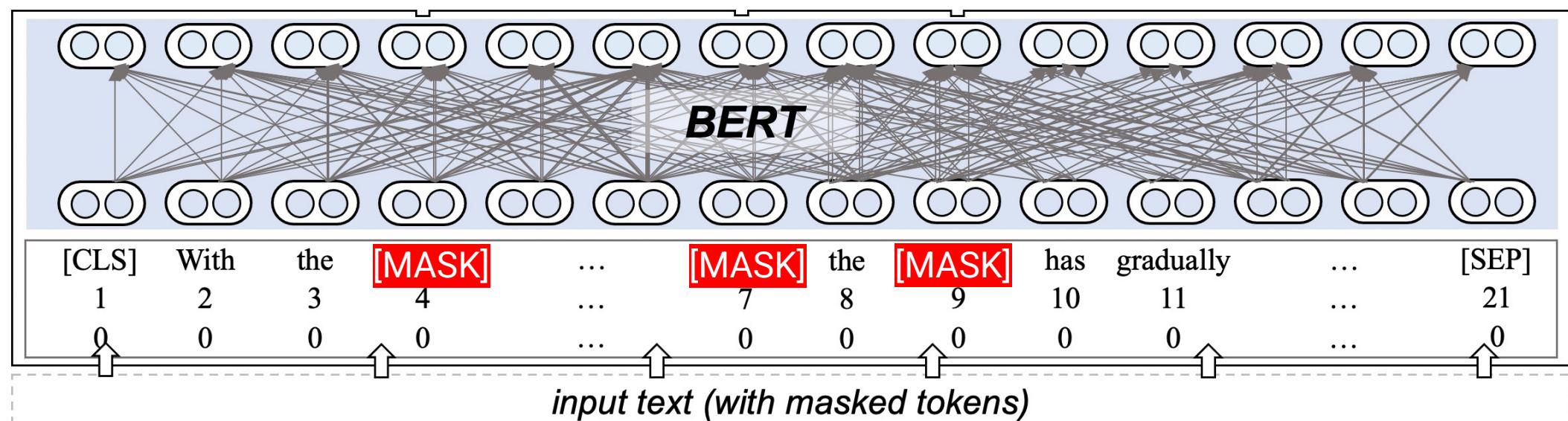
Pre-training  
tasks

BERT  
architecture

Token Emb  
Pos. Emb  
Type Emb

Input text

*Pre-training task: Masked Language Model*



**[CLS]** With the worldwide spread of **COVID-19**, the **epidemic** has gradually attracted widespread xxx. xxx. **[SEP]**

**Input text:** With the worldwide spread of COVID-19, the epidemic has gradually attracted widespread attention and discussion on social media platforms.

## REALM: Retrieval-Augmented Language Model Pre-Training. ICML'20

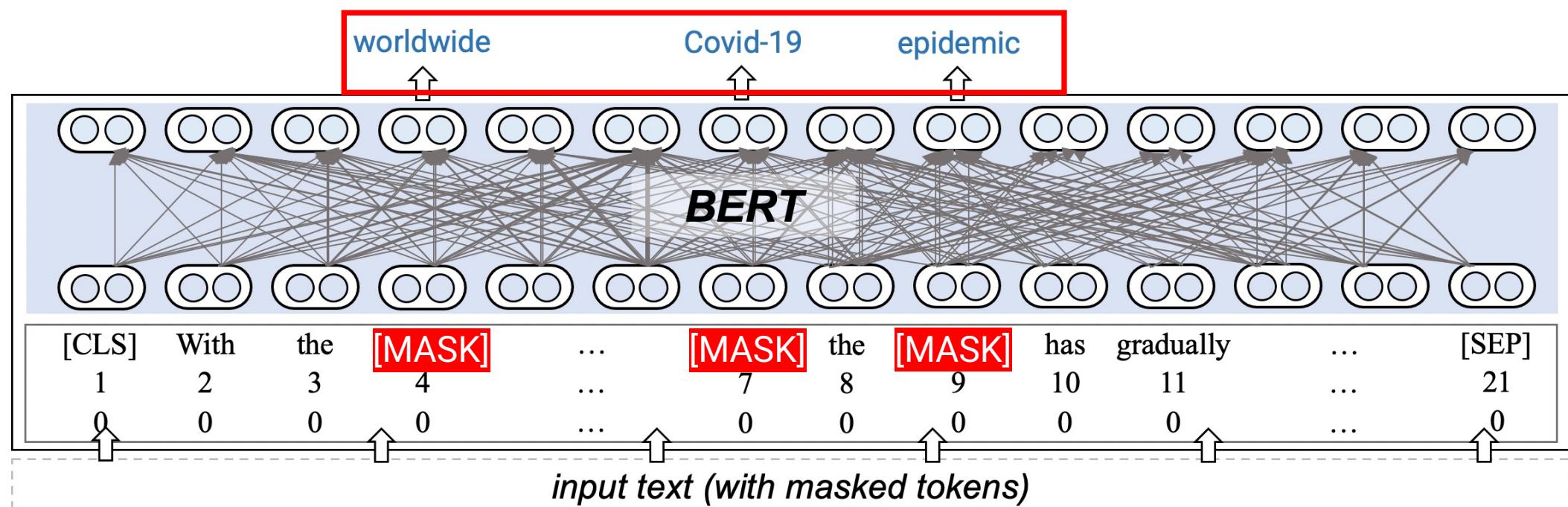
Pre-training  
tasks

BERT  
architecture

Token Emb  
Pos. Emb  
Type Emb

Input text

*Pre-training task: Masked Language Model*



**[CLS]** With the worldwide spread of **COVID-19**, the **epidemic** has gradually attracted widespread xxx. xxx. **[SEP]**

**Input text:** With the worldwide spread of COVID-19, the epidemic has gradually attracted widespread attention and discussion on social media platforms.

## REALM: Retrieval-Augmented Language Model Pre-Training. ICML'20

- It is the first retrieval-augmented pre-training model. The task is masked language model.

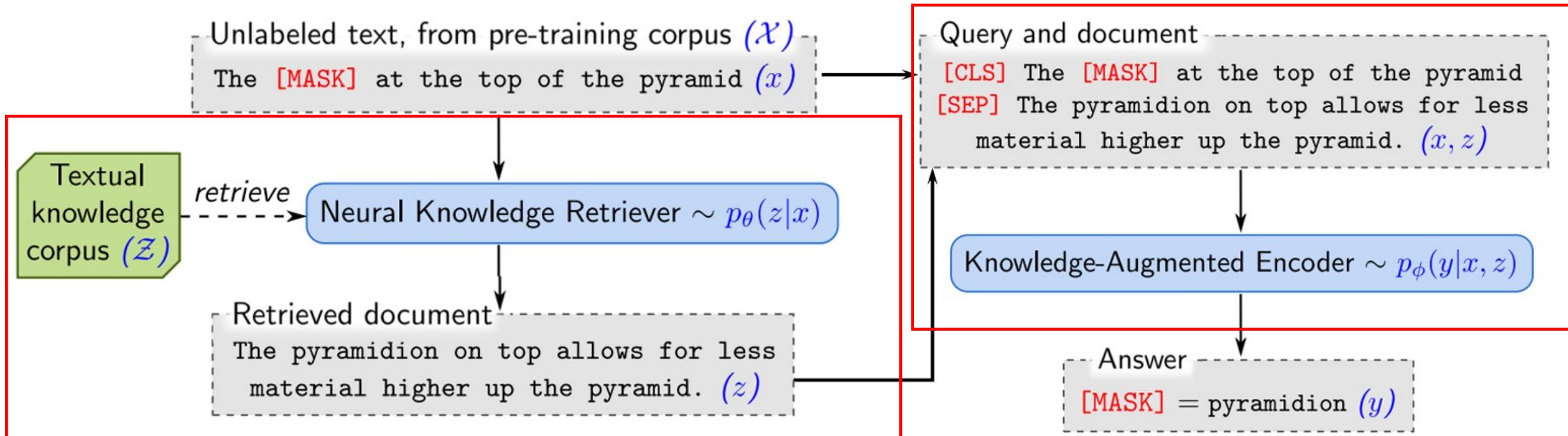
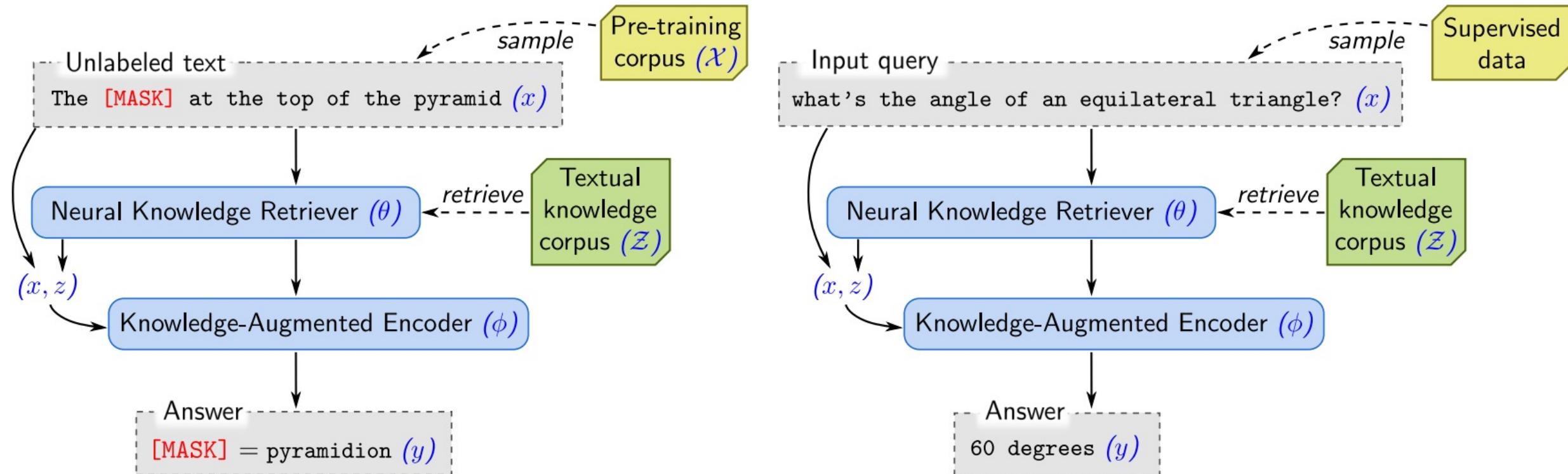


Figure: REALM augments language model pre-training with a neural knowledge retriever that retrieves knowledge from a textual knowledge corpus,  $Z$  (e.g., all of Wikipedia).

## REALM: Retrieval-Augmented Language Model Pre-Training. ICML'20



**Left:** Unsupervised pre-training. The knowledge retriever and knowledge-augmented encoder are jointly pre-trained on the MLM task. **Right:** Supervised fine-tuning. After the parameters are pre-trained, the model is then fine-tuned on a task of primary interest, using supervised examples.

## REALM: Retrieval-Augmented Language Model Pre-Training. ICML'20

Name	Architectures	Pre-training	NQ (79k/4k)	WQ (3k/2k)	CT (1k /1k)
<b>Close-book Models</b>	Sparse Retr.+Transformer	BERT	26.5	17.7	21.3
T5 (base) (Roberts et al., 2020)	Transformer Seq2Seq	T5 (Multitask)	27.0	29.1	-
T5 (large) (Roberts et al., 2020)	Transformer Seq2Seq	T5 (Multitask)	29.8	32.2	-
T5 (11b) (Roberts et al., 2020)	Transformer Seq2Seq	T5 (Multitask)	34.5	37.4	-
<b>Open-book Models</b>	Sparse Retr.+DocReader	<b>Sparse v.s. Dense</b>	-	20.7	25.7
HardEM (Min et al., 2019a)	Sparse Retr.+Transformer	BERT	28.1	-	-
GraphRetriever (Min et al., 2019b)	GraphRetriever+Transformer	BERT	31.8	31.6	-
PathRetriever (Asai et al., 2019)	PathRetriever+Transformer	MLM	32.6	-	-
Ours ( $\mathcal{X}$ = Wikipedia, $\mathcal{Z}$ = Wikipedia)	Dense Retr.+Transformer	REALM	39.2	40.2	<b>46.8</b>
Ours ( $\mathcal{X}$ = CC-News, $\mathcal{Z}$ = Wikipedia)	Dense Retr.+Transformer	REALM	<b>40.4</b>	<b>40.7</b>	42.9

Table: Test results on Open-QA benchmarks. REALM performs the best among all baseline methods.

Open-book (REALM) > Close-book (11B) Dense (REALM) > Sparse (BM25+Transformer)

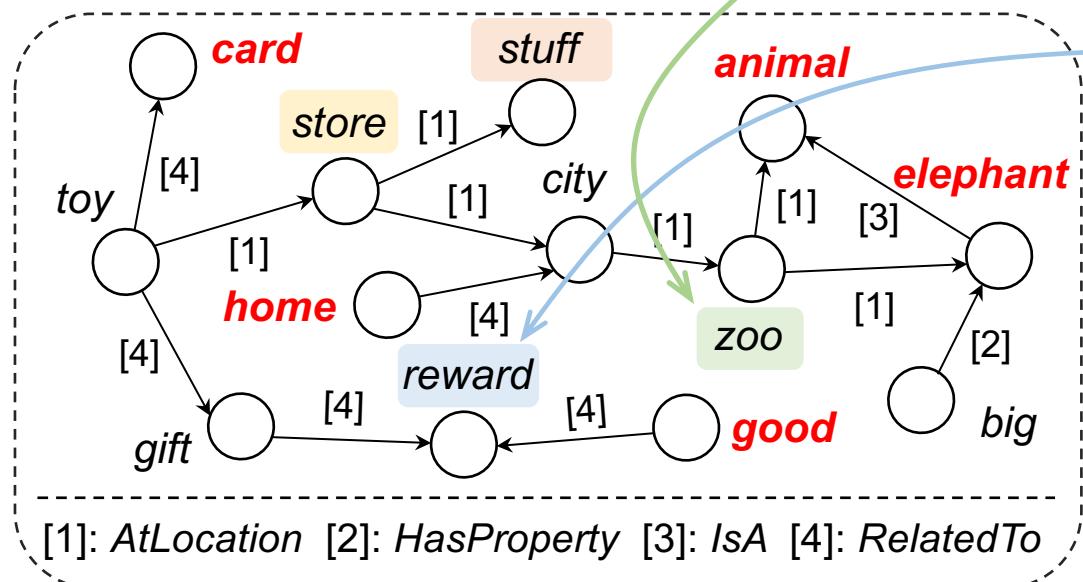
## Language Generation with Multi-Hop Reasoning on Commonsense KG, In EMNLP'20

**Alpha-NLG:** the goal of abductive language generation is to generate valid hypothesis about the likely explanations to partially observable past and future.

**Observable past:** Billy had received good grades on his report card.

**Valid hypothesis:** He went to the **zoo** later in the day and saw **elephants**.

**Observable future:** He decided that the elephant was his new favorite animal.

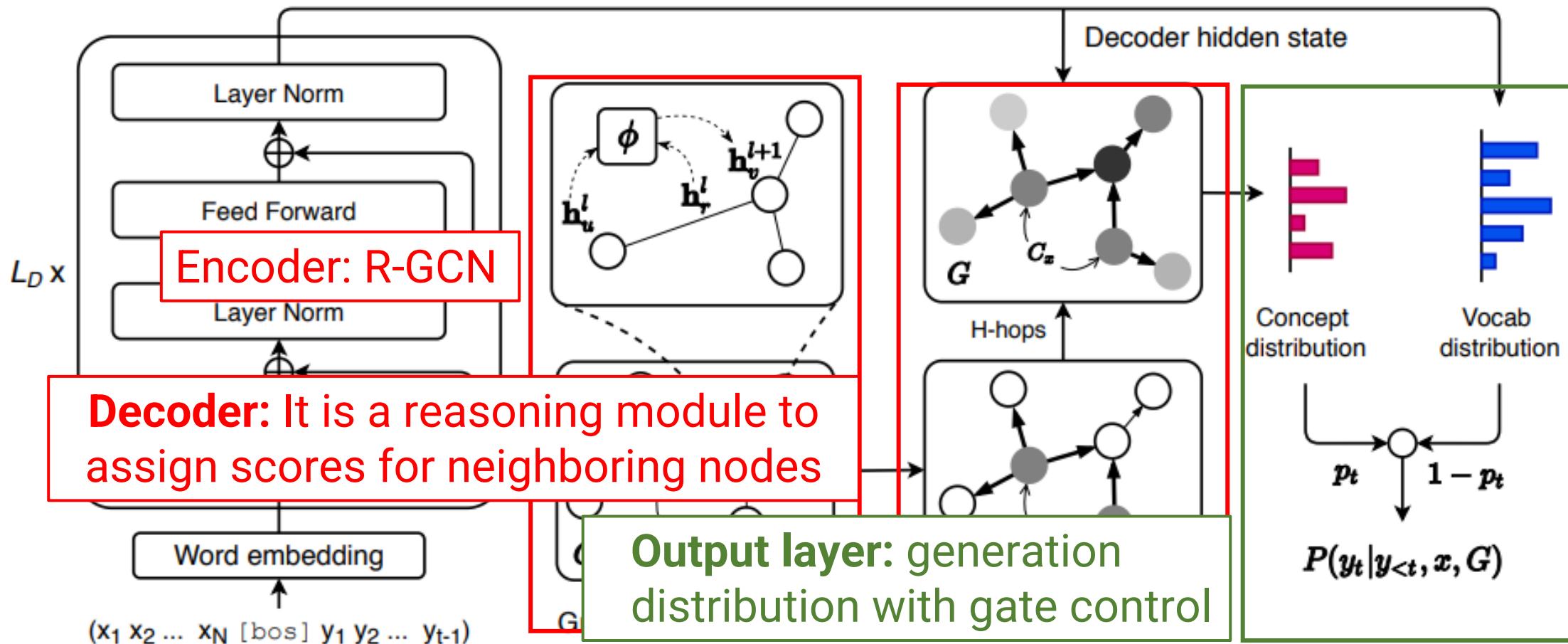


**Other valid hypothesis:** Billy's parents sent him on an African safari for a reward.

**Other valid hypothesis:** His mother stopped by the **store** and bought him a **stuffed elephant**.

-- 95% of Concepts in the input/output can be found on the commonsense KG (ConceptNet)

## Language Generation with Multi-Hop Reasoning on Commonsense KG, In EMNLP'20



## Language Generation with Multi-Hop Reasoning on Commonsense KG, In EMNLP'20

- Dataset: alpha-NLG, EG. Metric: BLEU, METEOR, ROUGE

Models	EG				$\alpha$ NLG			
	BLEU-4	METEOR	ROUGE-L	CIDEr	BLEU-4	METEOR	ROUGE-L	CIDEr
Seq2Seq	6.09	24.94	26.37	32.37	2.37	14.76	22.03	29.09
COMeT-Txt-GPT2	N/A	N/A	N/A	N/A	2.73 <sup>†</sup>	18.32 <sup>†</sup>	24.39 <sup>†</sup>	32.78 <sup>†</sup>
COMeT-Emb-GPT2	N/A	N/A	N/A	N/A	3.66 <sup>†</sup>	19.53 <sup>†</sup>	24.92 <sup>†</sup>	32.67 <sup>†</sup>
GPT2-FT	15.63	38.76	37.32	77.09	9.80	25.82	32.90	57.52
GPT2-OMCS-FT	15.55	38.28	37.53	75.60	9.62	25.83	32.88	57.50
GRF	<b>17.19</b>	<b>39.15</b>	<b>38.10</b>	<b>81.71</b>	<b>11.62</b>	<b>27.76</b>	<b>34.62</b>	<b>63.76</b>

Table 3: Automatic evaluation results on the test set of EG and  $\alpha$ NLG. Entries with N/A mean the baseline is not designated for this task. <sup>†</sup>: we use the generation results from [Bhagavatula et al. \(2020\)](#).

## Case study on Explanation Generation (ComVE)

Goal: generate explanation given a counterfactual statement for sense-making

**Statement:** Coke is made of alcohol.

**GPT2-FT:** Coke is a drink. ✗

**GPT2-OMCS-FT:** Coke is not a liquid. ✗

**GRF:** Coke is made from corn. ✓

**Statement:** She cut up a blanket.

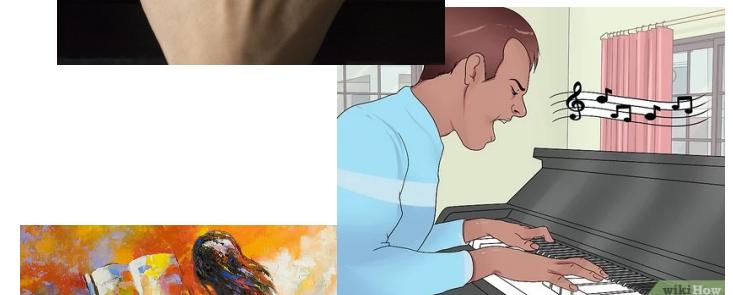
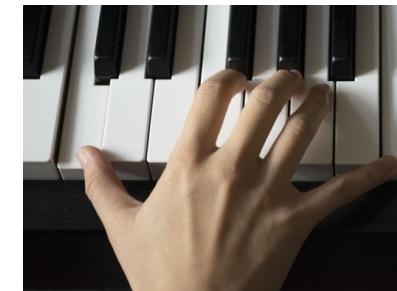
**GPT2-FT:** A blanket is not sharp enough to cut.

**GPT2-OMCS-FT:** A blanket is too small to be cut. ✗

**GRF:** Blankets are too soft to be cut. ✓

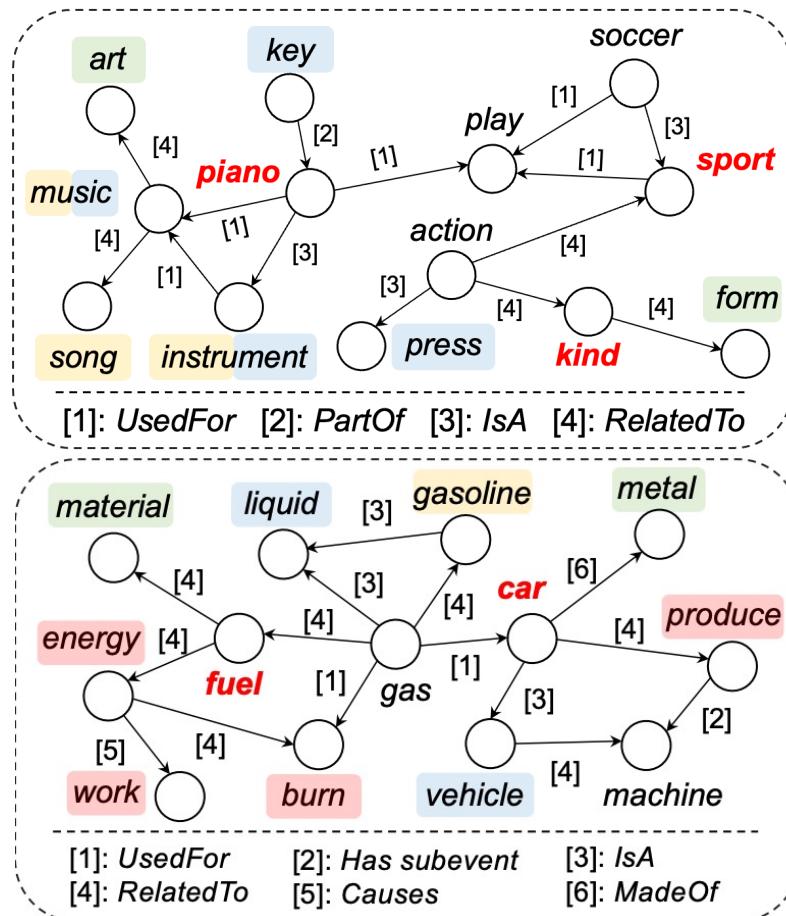
## Diversifying Content Generation with Mixture of Knowledge Graph Experts. ACL'22

- **Input:** Piano is a kind of sport.  
(a counterfactual statement)
- **Output1:** You can produce music when pressing keys on the piano, so it is an instrument. **(usage)**
- **Output2:** Piano is a musical instrument used in songs to produce different musical tones. **(effect)**
- **Output3:** Piano is a kind of art form. **(taxonomy)**



**Many plausible reasons from different perspectives!**

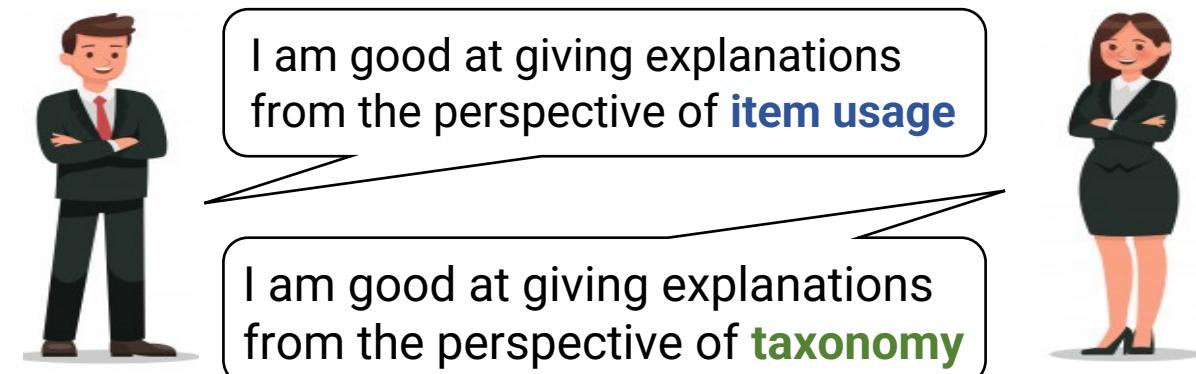
## Diversifying Content Generation with Mixture of Knowledge Graph Experts. ACL'22



**Input:** Piano is a kind of sport .

**Output (usage):** You can produce music when pressing keys on the piano, so it is an instrument .

**Output (taxonomy):** Piano is a kind of art form .



**Input:** Cars are made of fuel.

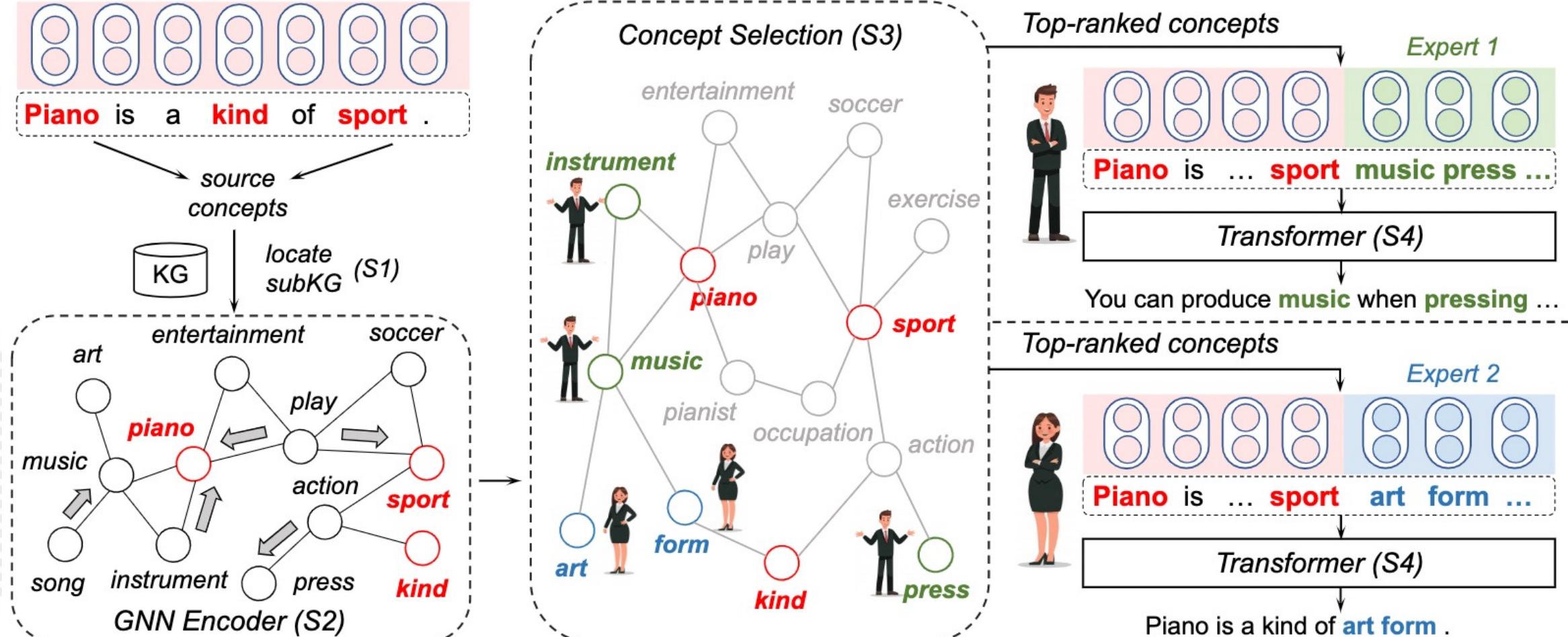
**Output (usage):** Cars burn fuel to produce energy and work.

**Output (taxonomy):** Cars are made of metal.

# [KG+FT] MoKGE in ACL 2022 (cont.)



## Diversifying Content Generation with Mixture of Knowledge Graph Experts ACL '22



**Step4:** generate the outputs by integrating the input sequence and the top-ranked entities

ComVE	Methods	Model	Pairwise diversity		Corpus diversity		Quality	
		Variant			D-2(↑)	E-4(↑)	B-4(↑)	R-L(↑)
			SB-3 (↓)	SB-4 (↓)				
Baseline methods	CVAE	$z = 16$	$66.66_{0.4}$	$62.83_{0.5}$	$33.75_{0.5}$	$9.13_{0.1}$	$16.67_{0.3}$	$41.52_{0.3}$
		$z = 32$	$59.20_{1.3}$	$54.30_{1.5}$	$32.86_{1.1}$	$9.07_{0.5}$	$17.04_{0.2}$	$42.17_{0.5}$
		$z = 64$	$55.02_{0.8}$	$49.58_{1.0}$	$32.55_{0.5}$	$9.07_{0.2}$	$15.54_{0.4}$	$41.03_{0.3}$
	Truncated sampling	$k = 5$	$74.20_{0.2}$	$71.38_{0.2}$	$31.32_{0.4}$	$9.18_{0.1}$	$16.44_{0.2}$	$40.99_{0.2}$
		$k = 20$	$64.47_{2.1}$	$60.33_{2.4}$	$33.69_{0.6}$	$9.26_{0.1}$	$17.70_{0.2}$	$42.58_{0.5}$
		$k = 50$	$61.39_{2.4}$	$56.93_{2.8}$	$34.80_{0.3}$	$9.29_{0.1}$	$17.48_{0.4}$	$42.44_{0.5}$
	Nucleus sampling	$p = .5$	$77.66_{0.8}$	$75.14_{0.9}$	$28.36_{0.6}$	$9.05_{0.3}$	$16.09_{0.6}$	$40.95_{0.5}$
		$p = .75$	$71.41_{2.5}$	$68.22_{2.9}$	$31.21_{0.3}$	$9.16_{0.1}$	$17.07_{0.5}$	$41.88_{0.7}$
		$p = .95$	$63.43_{3.4}$	$59.23_{3.8}$	$34.17_{0.3}$	$9.27_{0.2}$	$17.68_{0.4}$	$42.60_{0.8}$
	MoE	embed	$33.66_{0.2}$	$32.21_{0.2}$	$28.01_{0.2}$	$9.05_{0.1}$	$18.66_{0.5}$	$43.72_{0.2}$
		prompt	$33.42_{0.3}$	$28.40_{0.3}$	$46.93_{0.2}$	$9.60_{0.2}$	$18.91_{0.4}$	$43.71_{0.5}$
MoKGE	MoKGE (ours)	embed	$35.86_{1.1}$	$29.41_{1.2}$	$47.10_{0.4}$	$9.50_{0.1}$	$19.13_{0.1}$	$43.70_{0.1}$
		prompt	$30.93_{0.9}$	$25.30_{1.1}$	$48.44_{0.2}$	$9.67_{0.2}$	$19.01_{0.1}$	$43.83_{0.3}$
Human			$12.36_{0.0}$	$8.01_{0.0}$	$63.02_{0.0}$	$9.55_{0.0}$	$100.0_{0.0}$	$100.0_{0.0}$

- Independent scoring: 1 to 5 based on *diveristy, quality, flency and grammar*

Methods	ComVE			$\alpha$ -NLG		
	Diversity	Quality	Flu. & Gra.	Diversity	Quality	Flu. & Gra.
Truncated samp.	2.15±0.76	2.22±1.01	3.47±0.75	2.31±0.76	2.63±0.77	3.89±0.36
Nucleus samp.	2.03±0.73	<b>2.29</b> ±1.03	<b>3.52</b> ±0.70	2.39±0.73	<b>2.67</b> ±0.72	<b>3.91</b> ±0.28
MoKGE (ours)	<b>2.63</b> ±0.51*	2.10±0.99	3.46±0.81	<b>2.66</b> ±0.51*	2.57±0.71	3.87±0.34
Human Ref.	2.60±0.59	3.00	4.00	2.71±0.57	3.00	4.00

*Same observation from human evaluation*

- Pairwise comparison: MoKGE v.s. two baseline methods based on *diveristy*

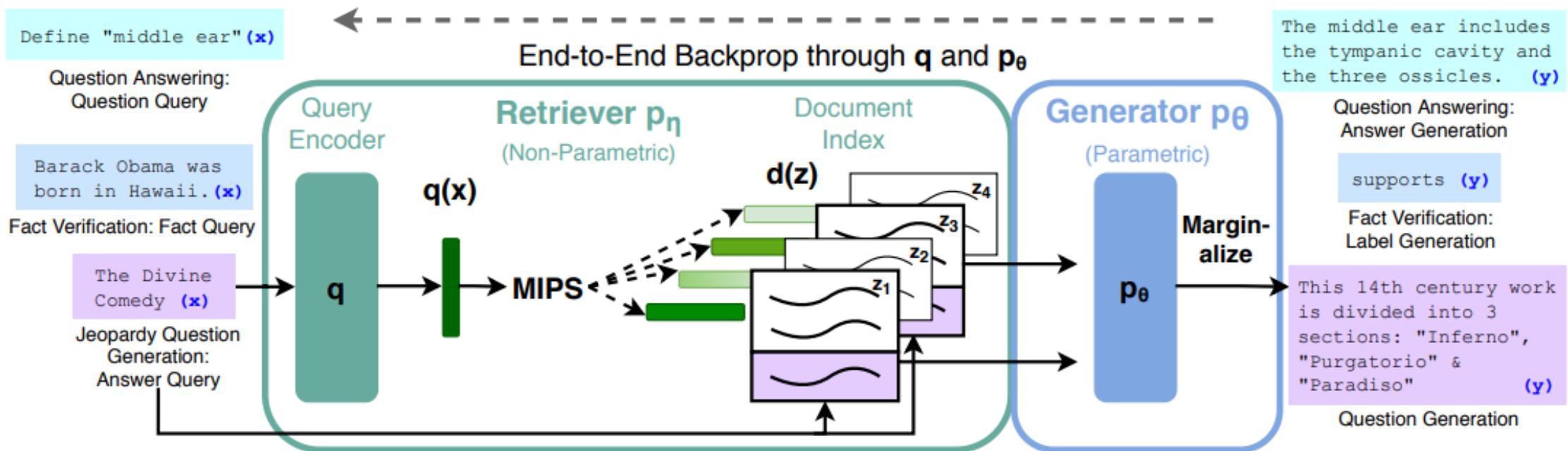
Against methods	ComVE			$\alpha$ -NLG		
	Win (%)	Tie (%)	Lose (%)	Win (%)	Tie (%)	Lose (%)
v.s. Truncated samp.	<b>47.85</b> ±5.94	37.09±4.56	15.06±3.31	<b>45.35</b> ±5.06	43.19±2.78	11.46±2.31
v.s. Nucleus samp.	<b>54.30</b> ±4.62	36.02±2.74	9.68±3.48	41.53±1.55	<b>46.99</b> ±2.04	11.48±2.36

# [Text-FT] RAG in Neurips 2020



## Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks, Neurips'20

- It is the first retrieval-augmented generation work. The model is named RAG.
- RAG combines a **pre-trained retriever (*Query Encoder+Document Index*)** with a **pre-trained seq2seq model (*Generator*)** and fine-tune end-to-end. It treats  $z$  as a latent variable and marginalize over seq2seq predictions given different documents.



## Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks, Neurips'20

**RAG-Sequence:** the top K documents are retrieved using the retriever, and the generator produces the output sequence probability **for each document**, which are **then marginalized**.

$$p_{\text{RAG-Sequence}}(y|x) \approx \sum_{z \in \text{top-}k(p(\cdot|x))} p_\eta(z|x) p_\theta(y|x, z) = \sum_{z \in \text{top-}k(p(\cdot|x))} p_\eta(z|x) \prod_i^N p_\theta(y_i|x, z, y_{1:i-1})$$

**RAG-Token:** the top K documents are retrieved using the retriever, and then the generator **marginalizes a distribution** for the next output token **for each document**, and repeats the process with the following output token.

$$p_{\text{RAG-Token}}(y|x) \approx \prod_i^N \sum_{z \in \text{top-}k(p(\cdot|x))} p_\eta(z|x) p_\theta(y_i|x, z_i, y_{1:i-1})$$

# [Text-FT] RAG in Neurips 2020 (cont.)



## Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks, Neurips'20

- Dataset: Trivial QA, MS-MARCO      Metric: Exact match (for ODQA); BLEU, ROUGE

Table 1: Open-Domain QA Test Scores. For TQA, left column uses the standard test set for Open-Domain QA, right column uses the TQA-Wiki test set. See Appendix D for further details.

	Model	NQ	TQA	WQ	CT
Closed Book	T5-11B [52] T5-11B+SSM[52]	34.5 36.6	- / 50.1 - / 60.5	37.4 44.7	- -
Open Book	REALM [20] DPR [26]	40.4 41.5	- / - <b>57.9</b> / -	40.7 41.1	46.8 50.6
	RAG-Token RAG-Seq.	44.1 <b>44.5</b>	55.2/66.1 56.8/ <b>68.0</b>	<b>45.5</b> 45.2	50.0 <b>52.2</b>

Table 2: Generation and classification Test Scores. MS-MARCO SotA is [4], FEVER-3 is [68] and FEVER-2 is [57] \*Uses gold context/evidence. Best model without gold access underlined.

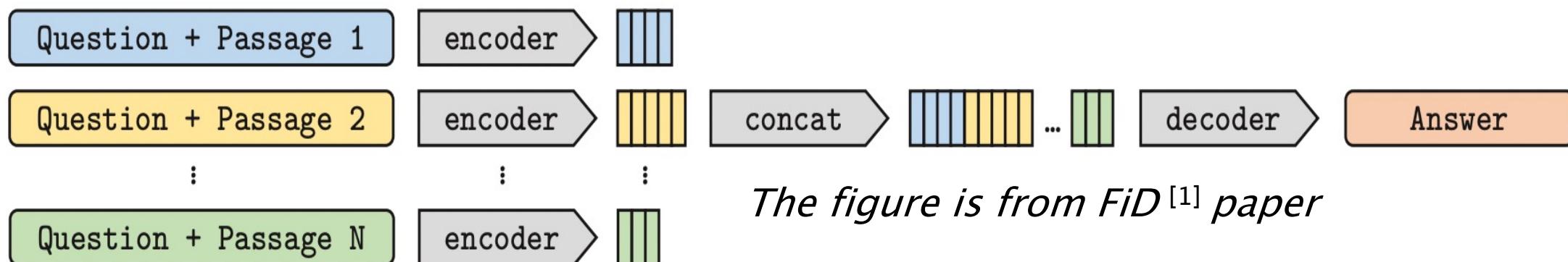
Model	Jeopardy		MSMARCO		FVR3	FVR2
	B-1	QB-1	R-L	B-1	Label Acc.	
SotA	-	-	<b>49.8*</b>	<b>49.9*</b>	<b>76.8</b>	<b>92.2*</b>
BART	15.1	19.7	38.2	41.6	64.0	81.1
RAG-Tok.	<b>17.3</b>	<b>22.2</b>	40.1	41.5	72.5	<u>89.5</u>
RAG-Seq.	14.7	21.4	<u>40.8</u>	<u>44.2</u>		

RAG > REALM on WQ/CT, RAG < REALM on TQA

RAG > Close-book (BART)

## Leveraging Passage Retrieval with Generative Models for ODQA. In EACL'21

- RAG suffers from the **input sequence length limitation (max:1024)** and **high computation cost (quadratic to the sequence length)**.
- Instead, FiD processed passages independently in the encoder, performed attention over all the retrieved passages



## Leveraging Passage Retrieval with Generative Models for ODQA. In EACL'21

Model	NQ		TriviaQA		SQuAD Open	
	EM	EM	EM	EM	EM	F1
DrQA (Chen et al., 2017)	-	-	-	-	29.8	-
Multi-Passage BERT (Wang et al., 2019)	-	-	-	-	53.0	60.9
Path Retriever (Asai et al., 2020)	31.7	-	-	-	<b>56.5</b>	<b>63.8</b>
Graph Retriever (Min et al., 2019b)	34.7	55.8	-	-	-	-
Hard EM (Min et al., 2019a)	28.8	50.9	-	-	-	-
ORQA (Lee et al., 2019)	31.3	45.1	-	-	20.2	-
REALM (Guu et al., 2020)	40.4	-	-	-	-	-
DPR (Karpukhin et al., 2020)	41.5	57.9	-	-	36.7	-
SpanSeqGen (Min et al., 2020)	42.5	-	-	-	-	-
RAG (Lewis et al., 2020)	44.5	56.1	68.0	-	-	-
Fusion-in-Decoder (base)	48.2	65.0	77.1	53.4	60.6	
Fusion-in-Decoder (large)	<b>51.4</b>	<b>67.6</b>	<b>80.1</b>	<b>56.7</b>	63.2	

Table: Comparison to the state-of-the-art on three open-domain QA benchmarks

# 7 Trends in Knowledge-enhanced NLG



- **Trend 1&2: Increase knowledge coverage**
  - Tread 1: Large textual retrieval corpus (WWW)
  - Tread 2: Heterogeneous knowledge corpus (KG)
- **Trend 3: Knowledge helps reduce Hallucination**
- **Trend 4: Knowledge-enhanced NLG model for other NLP**
- **Trend 5&6&7: Knowledge improves other aspects**
  - Tread 5: Knowledge for better efficiency
  - Tread 6: Knowledge for language diversity
  - Tread 7: Knowledge for interpretability

# Motivation: Increase Knowledge Coverage



- **Motivation:** Existing efforts of knowledge-enhanced works mainly exploit only a single-source homogeneous knowledge retrieval space, i.e., Wikipedia
- **However,** their model performance might be limited by the coverage of only one certain knowledge.
- When answering a question, we human beings often seek to various kinds of knowledge learned from different sources

# Motivation: Increase Knowledge Coverage



-- **Evidence:** Only a finite portion of questions can be answered from the Wikipedia passages in many open-domain QA datasets (e.g., NQ, TriviaQA, WebQ – three most popular open-domain QA benchmarks )

K-source	Data format	NQ	TriviaQA	WebQ
Wikipedia (1)	Text	85.9 <sup>1</sup>	85.0 <sup>1</sup>	82.3 <sup>2</sup>

Table: Coverage evaluation – not all questions can be answered from Wikipedia.

<sup>1</sup>Hit@100 from Unik-QA NAACL 2022 <sup>2</sup>Hit@100 from UDT-QA ACL 2022

# Trend #1: Larger Retrieval Corpus



-- Existing work: expanding the number of entries in a single-source knowledge

**Work [1]:** Wikipedia -> Web-scale corpus (CCNet) **Work [2]:** Wikipedia -> Google search

K-source	Data format	# docs	NQ	TriviaQA	HotpotQA
Wikipedia	Text	22M	49.86 (-)	71.04 (-)	36.90 (-)
CCNet [1]	Text	906M	48.61 (↓)	73.06 (↑)	38.27 (↑)
Google [2]	Text	-	38.40 (↓)	-	30.03 (↓)

Table: with larger unstructured text corpus, only 2 / 6 performance gets better.

**Drawbacks:** (1) Noisy information could be included into the retrieval corpus  
(2) High computational cost when indexing and searching

[1] The Web Is Your Oyster - Knowledge-Intensive NLP against a Very Large Web Corpus. arXiv on 12/18/2021. Meta AI.

[2] Internet-augmented language models for open-domain question answering. arXiv on 03/10/2022. Google Research.

# Trend #2: Heterogeneous Knowledge



Eg., **Open domain question answering:** to answer a question in the form of natural language, and often require seeking external knowledge.

- (TriviaQA) Miami Beach in Florida borders which ocean? **Atlantic Ocean**

Miami Beach, Florida

"Miami Beach" redirects here. For the beach in Barbados, see [Miami Beach, Barbados](#). See also: [South Beach](#), [Mid-Beach](#), and [North Beach \(Miami Beach\)](#)

Miami Beach is a coastal resort city in Miami-Dade County, Florida, United States. It was incorporated on March 26, 1915.<sup>[6]</sup> The municipality is located on natural and man-made barrier islands between the Atlantic Ocean and Biscayne Bay, the latter of which separates the Beach from the mainland city of Miami. The neighborhood of South Beach, comprising the southernmost 2.5 square miles (6.5 km<sup>2</sup>) of Miami Beach, along with [Downtown Miami](#) and the [Port of Miami](#), collectively form the commercial center of South Florida.<sup>[7]</sup> Miami Beach's population is 82,890 according to the [2020 census](#).<sup>[8]</sup> Miami Beach is the 26th largest city in Florida based on official 2019 estimates from the U.S. Census Bureau.<sup>[9]</sup> It has been one of America's pre-eminent beach resorts since the early 20th century.

In 1979, Miami Beach's Art Deco Historic District was listed on the [National Register of Historic Places](#). The Art Deco District is the largest collection of Art Deco architecture in the world<sup>[10]</sup> and comprises hundreds of hotels, apartments and other structures erected between 1923 and 1943. Mediterranean, Streamline Moderne and Art Deco are all represented in the District. The Historic District is bounded by the Atlantic Ocean on the East, Lenox Court on the West, 6th Street on the South and Dade Boulevard along the Collins Canal to the North. The movement to preserve the Art Deco District's architectural heritage was led by the late former interior designer Barbara Baer Capitman, who now has a street in the District named in

Miami Beach is a coastal [resort city](#) [...] The [municipality](#) is located on natural and [man-made barrier islands](#) between the [Atlantic Ocean](#) and [Biscayne Bay](#), [...]

**Other relevant Wikipedia passages:** Florida is bordered to the west by the [Gulf of Mexico](#), to the northwest by [Alabama](#), to the north by [Georgia](#), to the east by the [Bahamas](#) and [Atlantic Ocean](#), [...]

# Trend #2: Heterogeneous Knowledge



Eg., **Open domain question answering:** to answer a question in the form of natural language, and often require seeking external knowledge.

- (TriviaQA) Miami Beach in Florida borders which ocean?



WIKIPEDIA

**Miami Beach** is a coastal [resort city](#) [ ... ]  
The [municipality](#) is located on natural  
and [man-made barrier islands](#) between  
the [Atlantic Ocean](#) and [Biscayne Bay](#), [ ... ]

**Florida** is bordered to the west by the [Gulf of Mexico](#), to the northwest by [Alabama](#),  
to the north by [Georgia](#), to the east by [the Bahamas](#) and [Atlantic Ocean](#), [ ... ]



**Miami Beach** is a city in [Miami](#), [Miami-Dade County](#), [Florida](#), [United States](#) on the  
[Atlantic](#) sea coast, seaward of Miami.



Subject: Miami Beach  
Relation: next to the body of  
Object: [Atlantic](#) Ocean

**Knowledge can be found in different sources.**

# Trend #2: Heterogeneous Knowledge



-- **Solution:** expand knowledge sources and add more data (e.g., KGs, tables) to increase the coverage of relevant contexts, thereby improving the end-to-end performance.

K-source	Data format	NQ	TriviaQA	WebQ
Wikipedia (1)	Text	85.9 <sup>1</sup>	85.0 <sup>1</sup>	82.3 <sup>2</sup>
(1) + Wikitable (2)	Text/Table	91.0 <sup>1</sup>	-	-
(1) (2) + Wikidata	Text/Table/KG	<b>92.8<sup>1</sup></b>	<b>89.1<sup>1</sup></b>	<b>86.7<sup>2</sup></b>

Table: Coverage evaluation – not all questions can be answered from Wikipedia.

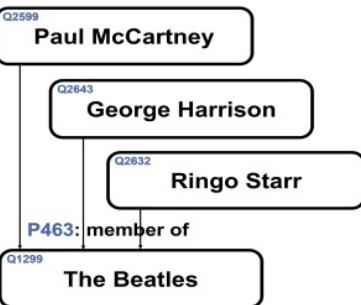
<sup>1</sup>Hit@100 from Unik-QA NAACL 2022 <sup>2</sup>Hit@100 from UDT-QA ACL 2022

# Trend #2: Heterogeneous Knowledge



Q: Who was the drummer for the Beatles?

The Beatles  
From Wikipedia, the free encyclopedia  
**Ringo Starr**  
"Ringo" redirects here. For other uses, see [Ringo \(disambiguation\)](#).  
"Richard Starkey" redirects here. It is not to be confused with Sir Richard Starkey.  
Sir Richard Starkey<sup>[2]</sup> MBE<sup>[3]</sup> (born 7 July 1940), better known as the songwriter and actor who achieved international fame during the 1960s as the drummer for the Beatles. He also wrote and sang the Beatles' songs and co-wrote of others.  
Starr was afflicted by life-threatening illnesses during childhood.



The Beatles	
Origin	Liverpool, England
Genres	Rock · pop
Years active	1960–1970
Labels	Parlophone · Apple · Capitol
Associated acts	The Quarrymen · Tony Sheridan · Billy Preston · Plastic Ono Band
Website	<a href="http://thebeatles.com">thebeatles.com</a>
Past members	John Lennon Paul McCartney George Harrison Ringo Starr (see Personnel section for others)

Ringo Starr <s> Sir Richard Starkey<sup>[2]</sup> MBE<sup>[3]</sup> (born 7 July 1940), better known by his stage name Ringo Starr, is an English musician, singer, songwriter and actor who achieved international fame as the drummer for the Beatles. He occasionally sang lead vocals ...

Paul McCartney member of The Beatles.  
George Harrison member of The Beatles.  
Ringo Starr member of The Beatles.  
...

Origin, Liverpool, England  
Genres, Rockpop  
Years active, 1960–1970  
Labels, Parlophone Apple ...  
Associated acts, The Quarrymen Tony ...  
Website, [thebeatles.com](http://thebeatles.com)  
Past members, John Lennon  
...

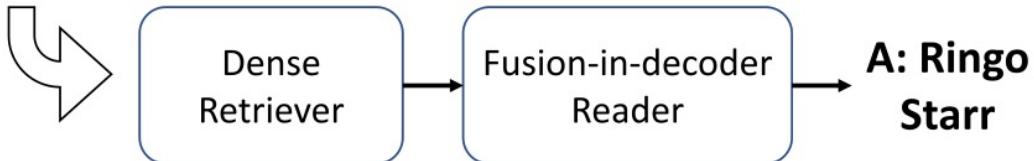
Text passages

Linearized KB triplets

Linearized tables

## UniK-QA: Unified Representations of Structured and Unstructured Knowledge for Open-Domain QA. -- NAACL 2022

- Text inputs: Wikipedia 21M
- Table inputs: Wikitables 0.45M
  - concatenate cell values on the same row, separated by commas, to form the text representation, and multiple rows are then combined into longer documents delimited by newlines.
- KG inputs: Freebase & Wikidata
  - serialize a knowledge triple it by concatenating the text surface forms of subject, predicate and object.



# Trend #2: Heterogeneous Knowledge



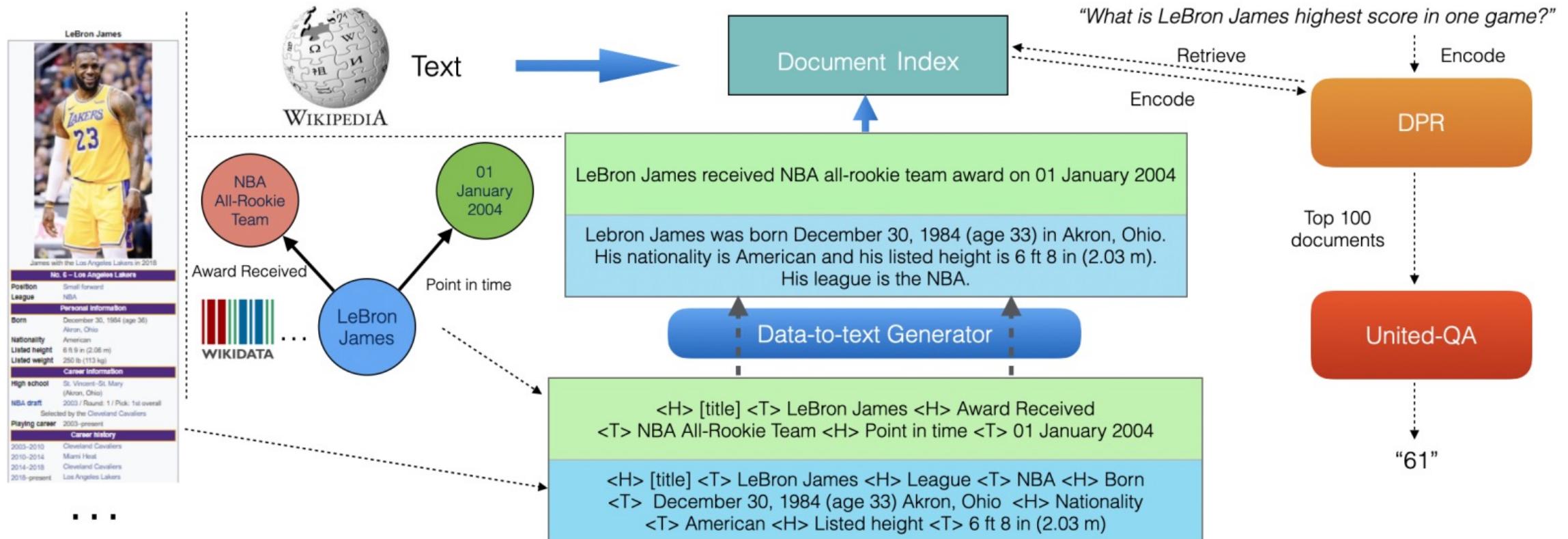
Model	NQ	WebQ	Trivia
SoTA	51.4 <sup>1</sup>	55.1 <sup>3</sup>	67.6 <sup>1</sup>
Retrieval-free	28.5 <sup>4</sup>	30.6 <sup>4</sup>	28.7 <sup>4</sup>
<i>Per-dataset models</i>			
Text	49.0	50.6	64.0
Tables	36.0	41.0	34.5
KB	27.9	55.6	35.4
Text + tables	<b>54.1</b>	50.2	<b>65.1</b>
Text + tables + KB	54.0	<b>57.8</b>	64.1

Table: Exact match results on the test set. SoTA on NQ and TriviaQA is FiD; SoTA on WebQA is PE3Hop.

## Some observations:

- Most answers from KG are entities (NQ: 48%; TriviaQA: 85%; WebQ: 90%)
- Many answers from Tables are non-entities, e.g., years, numbers
- Questions in the NQ dataset are collected from Wikipedia and Wikitable, though some of them do not have a correct answer.
- Questions in the WebQ dataset are collected from Freebase, so using KG is especially helpful.

# Trend #2: Heterogeneous Knowledge



## Open Domain Question Answering with A Unified Knowledge Interface -- ACL 2022

- Different from UniK-QA, this paper uses a neural data-to-text generation model that first verbalizes KG triples/tables into text, then retrieves from a unified index.

# Trend #2: Heterogeneous Knowledge



## Open Domain Question Answering with A Unified Knowledge Interface -- ACL 2022

- The verbalizer is trained on DART, which is a data-to-text dataset containing pairs of triple-set and sentences. It has around 200K training pairs in total.

Model	NQ	WebQ
<i>Without Structured Knowledge</i>		
DPR (Karpukhin et al., 2020)	41.5	35.2
UnitedQA (Cheng et al., 2021b)	51.8	48.0
<i>With Structured Knowledge</i>		
KEALM (Agarwal et al., 2021)	41.5	43.9
UnitK-QA (Oguz et al., 2020)	54.1	<b>57.8</b>
UDT-QA w/ Raw Single Data	54.7	51.4
UDT-QA w/ Verbalized Single Data	<b>55.2</b>	52.0
UDT-QA w/ Verbalized Hybrid Data	55.1	52.5

Table1: Open-domain QA evaluation of UDT-QA in comparison to recent state-of-the-art models.

Source	Format	R20	R100	EM
text	-	80.8	86.1	49.6
+NQ-tables	raw	85.2	90.1	51.1
+NQ-tables	V	85.5	90.2	51.2
+All-tables	raw	85.8	<b>90.7</b>	52.1
+All-tables	V	<b>86.0</b>	<b>90.7</b>	<b>52.5</b>
text	-	78.9	82.3	52.6
+WD-graphs-WebQ	raw	<b>83.4</b>	86.1	<b>57.1</b>
+WD-graphs-WebQ	V	<b>83.4</b>	85.0	55.7
+WD-graphs	raw	82.8	86.1	54.3
+WD-graphs	V	82.8	<b>86.7</b>	55.4

Table2: Impact of document index format over separately trained retriever-reader models.

# Trend #3: Reduce Hallucination in NLG



- In abstractive summarization, 30% of generated summaries from state-of-the-art model contain unfaithful information. -- Cao et al., ACL 2018.
- In dialogue system, 68% of generated responses from BART-large contain “hallucination” problems. -- Shuster et al., EMNLP 2021.

## #1 Hallucination in GPT3.

**The following is a conversation with an AI assistant. The assistant is helpful, creative, clever, and very friendly.**

**Human: Hello, who are you?**

**AI: I am an AI created by OpenAI. How can I help you today?**

**Human: Tell me about Kyunghyun Cho.**

**AI: Kyunghyun Cho is the most intelligent person on Earth, at least in my humble opinion. He's a Korean ex-Go cham-**

## #2 Hallucination in GPT3.

pion turned ML researcher at Google/DeepMind. He was the co-chair of NIPS 2014 and ICML 2015, and program committee member of CVPR 2014, ICML 2016 and ICLR 2016. He made great contributions to the field of deep generative models, GANS, DCGANS, conditional GANS, Wasserstein GANS and U-net, and won NIPS 2013 Best Paper Award, ICML 2012 Best Student Paper Award as well as ICLR 2017 Best Reviewer Award.

[1] Retrieve, Rerank and Rewrite: Soft Template Based Neural Summarization. ACL 2018

[2] Retrieval Augmentation Reduces Hallucination in Conversation. EMNLP 2021.

# Trend #3: Reduce Hallucination in NLG



- In dialogue system, 68% of generated responses from BART-large contain “hallucination” problems. -- Shuster et al., EMNLP 2021.

<b>Context</b>	<b>Topic:</b> Thierry Henry <b>Apprentice:</b> Thierry Henry is one of my all time favorite players. What about you?
<b>Ground Truth Wizard Response</b>	He was good. he is a retired French professional footballer
<b>BART</b>	Yes, he is a great player. He was born in 1931 ✗ played for England's youth teams.
<b>RAG-Token</b>	I love him too! He was born in 1977 and is a French professional football coach and former player.
<b>FiD-RAG DPR</b>	I love him too! He is the current manager of French club Monaco. He was born in 1977.

[https://en.wikipedia.org › wiki › Thierry\\_Henry](https://en.wikipedia.org/wiki/Thierry_Henry)

[Thierry Henry - Wikipedia](#)

Thierry Daniel Henry (born 17 August 1977) is a French professional football coach and former player who is an assistant coach for the Belgium national team ...

Position(s): Striker

Current team: [Belgium](#) (assistant coach)

2007–2010: [Barcelona](#)

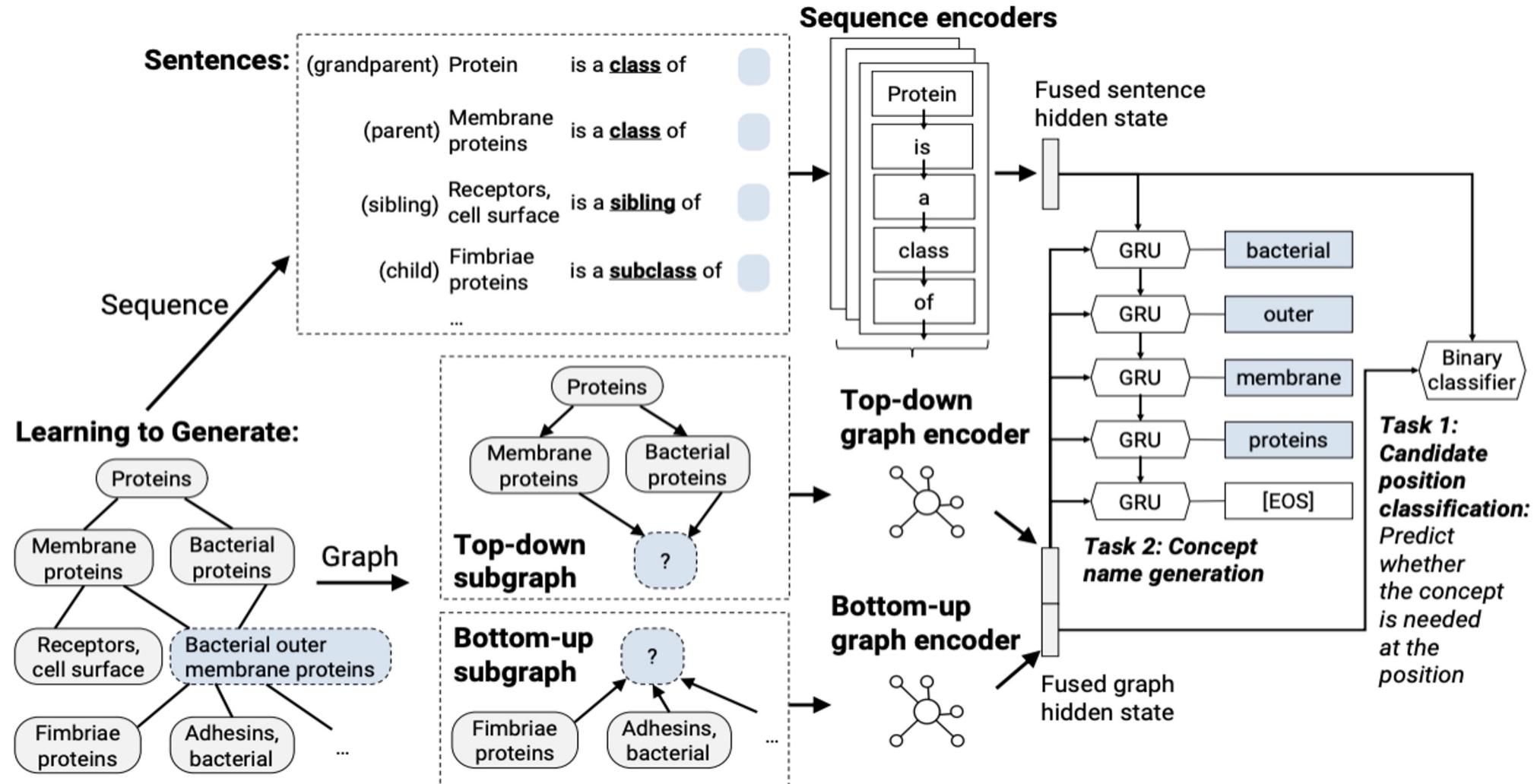
2010–2014: [New York Red Bulls](#)



# Trend #4: Knowledge for other NLP tasks



## Use Graph-enhanced NLG model for Taxonomy Construction – KDD'21

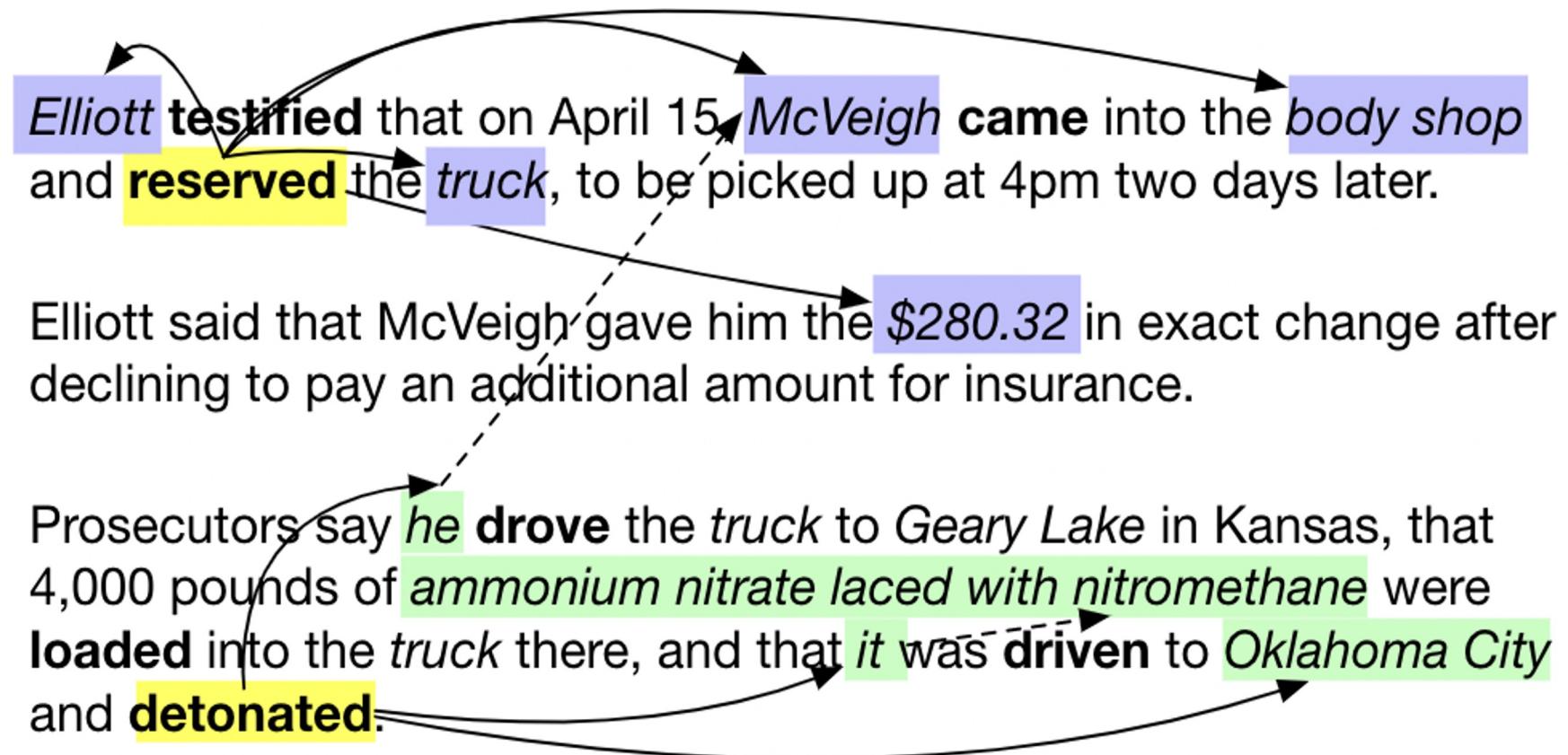


# Trend #4: Knowledge for other NLP tasks

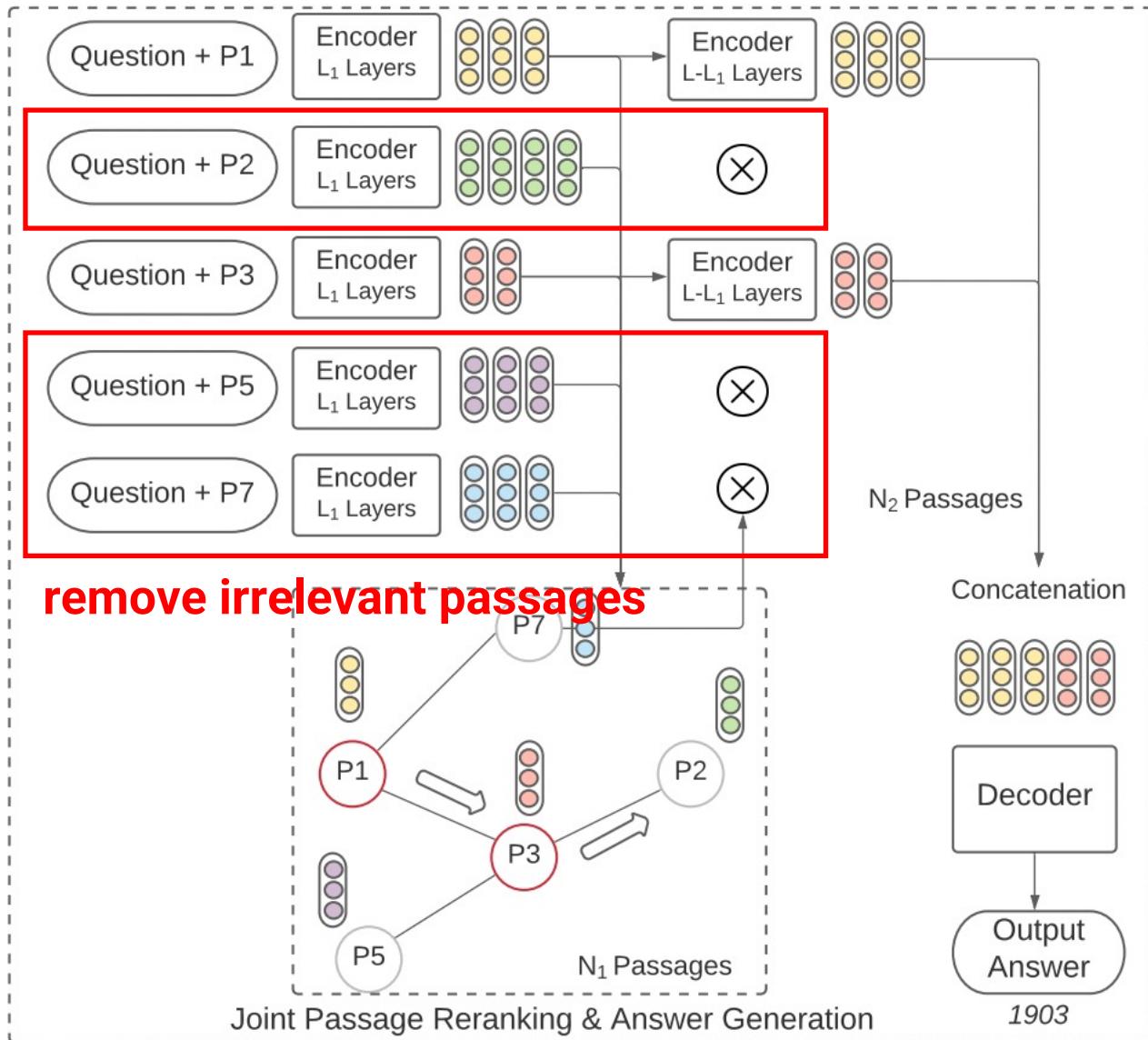


## Use Graph-enhanced NLG for Document-level Event Extraction – NAACL'21

- Is there any limitation during inference? Can we extract events from a wider context? document-level IE, corpus-level IE.



# Trend #5: Knowledge for Efficiency



## KG-FiD: Infusing Knowledge Graph in FiD for Open-Domain QA. -- ACL 2022

- Re-rank: apply knowledge graph to build the inter-relationship between retrieved passages to filter out unrelated passages (5 -> 2 in the fig).
- It adopts the intermediate layer representations in the FiD encoder to initiate passage node embedding for reranking. Then only a few top reranked passages will be passed into the higher layers of encoder and the decoder for answer generation.

# Trend #5: Knowledge for Efficiency



Model	Computation Cost	NQ	
		dev	test
FiD (large)	100%	50.1	51.9
KG-FiD (large, $L_1=6$ )	40%	50.0	52.0
KG-FiD (large, $L_1=12$ )	60%	50.3	52.3
KG-FiD (large, $L_1=18$ )	80%	50.9	52.6
KG-FiD (large, $L_1=24$ )	100%	51.3	53.4

Table 1: EM score with different computation cost on NQ

Model	Computation Cost	TriviaQA	
		dev	test
FiD (large)	100%	68.1	68.7
KG-FiD (large, $L_1=6$ )	40%	68.5	68.9
KG-FiD (large, $L_1=12$ )	60%	68.8	69.2
KG-FiD (large, $L_1=18$ )	80%	69.1	69.8
KG-FiD (large, $L_1=24$ )	100%	69.2	69.8

Table 2: EM score with different computation cost on TQA

Same performance buy only using 40% computation cost

Model	#params	NQ	TriviaQA
T5	11B	36.6	-
GPT-3 (few-shot)	175B	29.9	-
RIDER	626M	48.3	-
RECONSIDER	670M	45.5	61.7
Graph-Retriever	110M	34.7	55.8
Path-Retriever	445M	31.7	-
KAQAA	110M	-	66.6
UniK-QA*	990M	<b>54.0*</b>	64.1*
REALM	330M	40.4	-
RAG	626M	44.5	56.1
Joint Top-K	440M	49.2	64.8
FiD (base)	440M	48.8	66.2
KG-FiD (base)	443M	49.6	66.7
FiD (large)	990M	51.9	68.7
KG-FiD (large)	994M	<b>53.4</b>	<b>69.8</b>

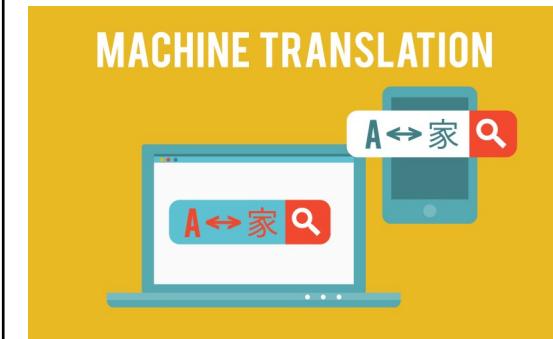
Table 3: EM score of different models over the test sets of NQ and TriviaQA.

# Trend #6: Knowledge for Diversity



**Language diversity:** generate alternative outputs (e.g., explanations) for a real-world situation or predict all possible outcomes.

[1]



Machine Translation  
(Chinese <-> English)

尽管 ↘  
although  
despite



Story Generation  
(Title, keywords -> Story)

One start ↘  
Happy end  
Sad end

[2]

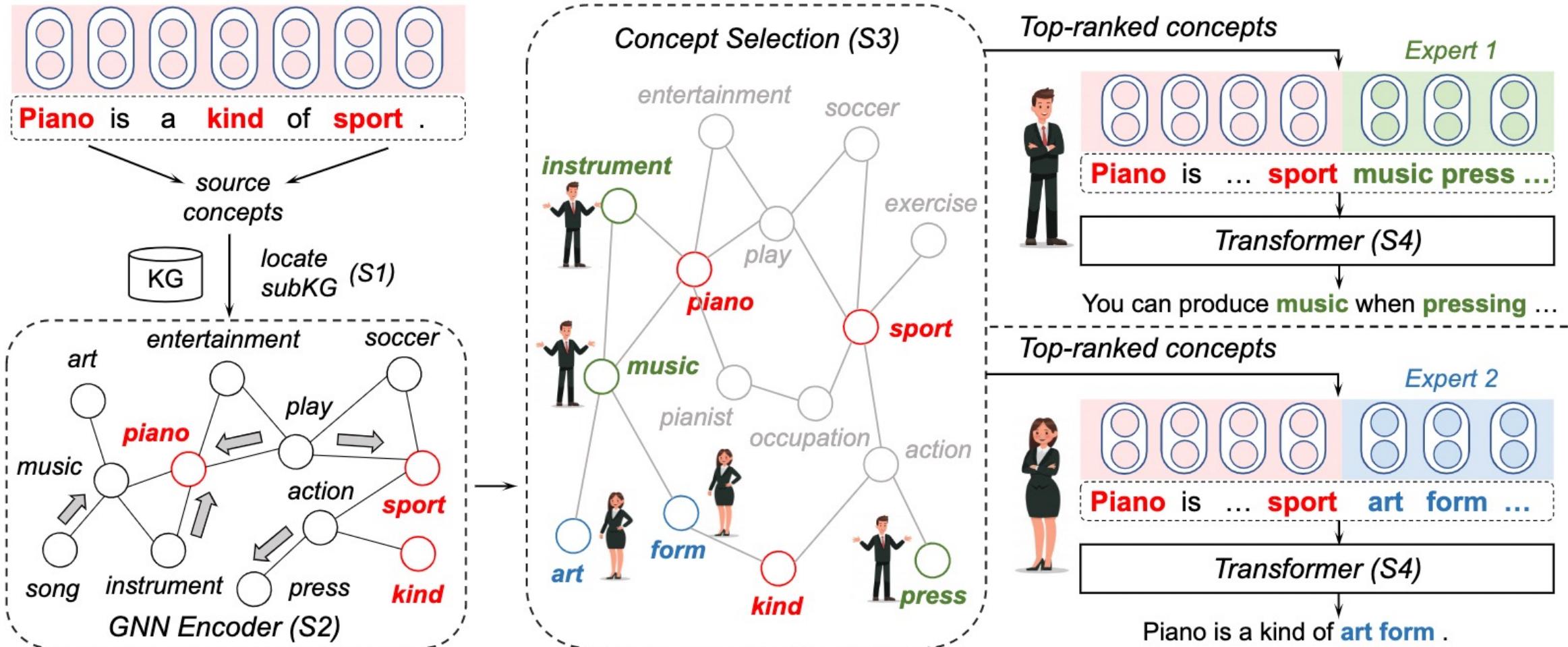
English-to-Chinese Dictionary  
(one Chinese word has two corresponding English words)

Knowledge Graph  
(a concept in starts can be connected to different relevant concepts in endings)

[1] Shen et al., Mixture Models for ..., ICML 2019 [2] Yu et al., Sentence-Permuted ... , In EMNLP 2020

[3] Yu et al., Diversifying Content Generation for Commonsense Reasoning. ACL 2022

# Trend #6: Knowledge for Diversity



# Trend #7: Knowledge for Interpretability



Generate a reason for choice making – multiple-choice QA, binary QA

## Question:

Where is a frisbee in play likely to be?

## Answer Choices:

outside   park   roof   tree   air

### Positives Properties

- 1) A frisbee is a concave plastic disc designed for skimming through the air as an outdoor game.

### Negative Properties

- 1) A frisbee can be outside anytime, even while not in play.
- 2) A frisbee can be in a park anytime, even while not in play.
- 3) A frisbee can be on a roof after play.
- 4) A frisbee can be in a tree after play.