WSDM 2023 Tutorial

Knowledge-Augmented Methods for Natural Language **Generation**

Wenhao Yu and Meng Jiang
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 …, 2022 - dl.acm.org

… **Wenhao Yu** and Dr. Meng Jiang's research is supported by National Science Foundation grants IIS-1849816, CCF-1901059, and IIS-2119531. Qingyun Wang and Dr. …

☆ Save Cite Cited by 87 Related articles All 9 versions

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

Reading List:

Knowledge-enriched Text Generation Survey, Tutorial and Reading

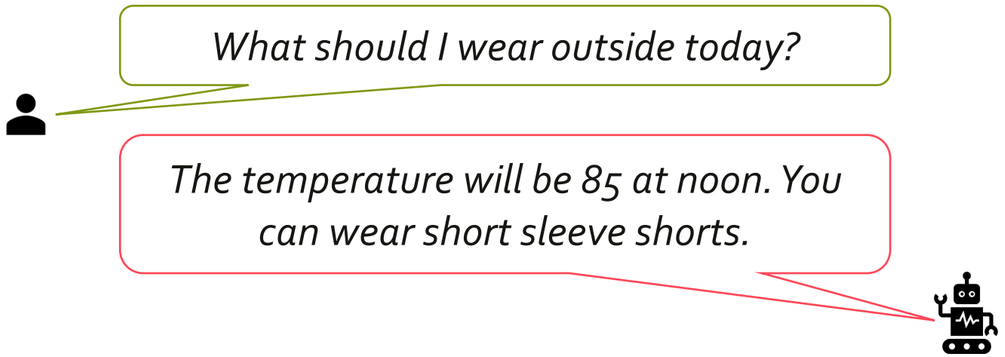
Status building PRs Welcome Unwatch 32 Fork 49 Starred 504

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.

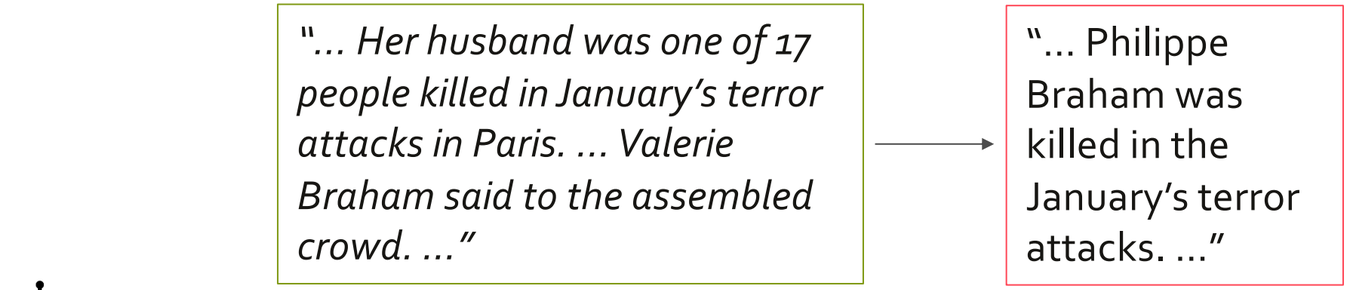
- Github: https://github.com/wyu97/KENLG-Reading

Text generation is everywhere in our life!

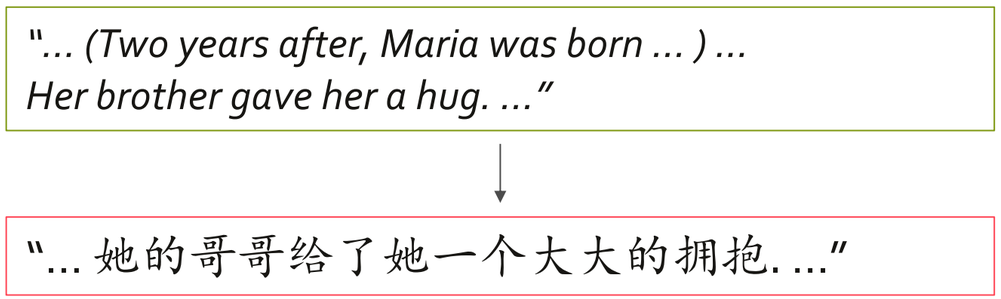
- Leveraging machine intelligence to make many things EASY and FAST



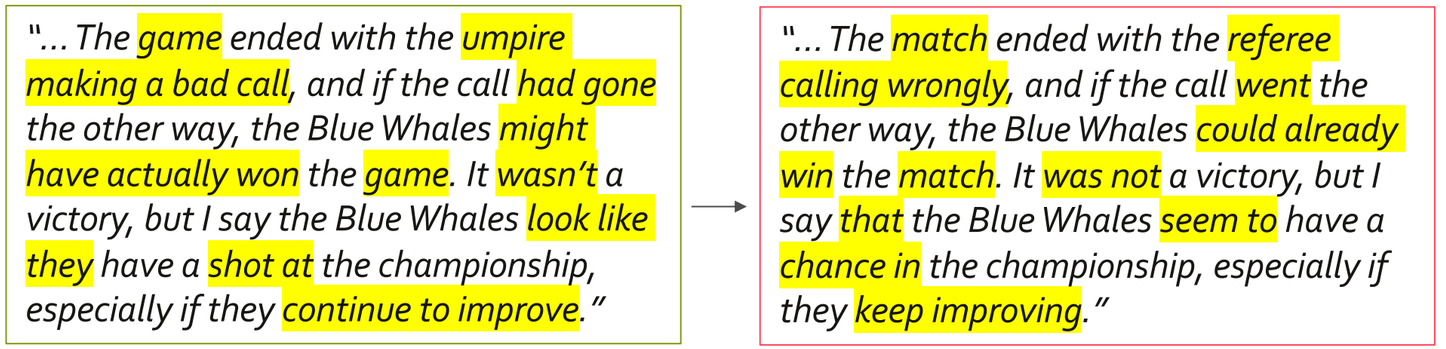
Dialog systems



Summarization



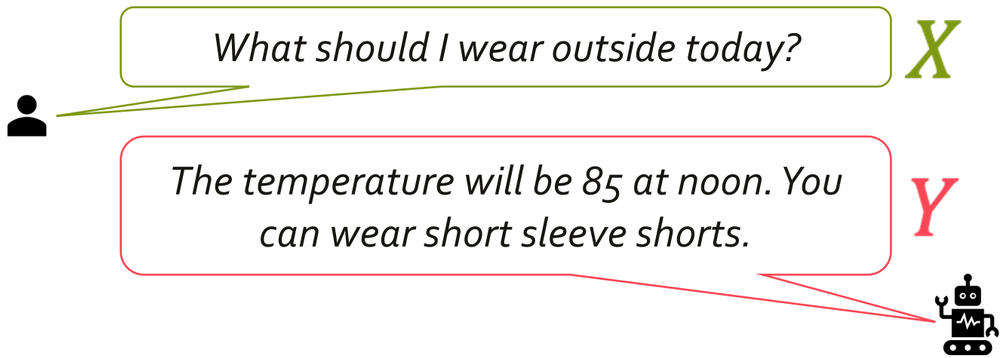
Machine Translation



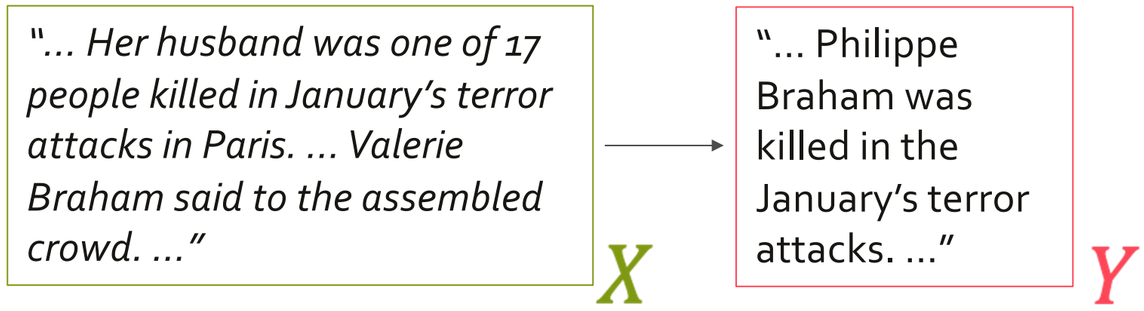
Paraphrasing

Text generation is everywhere in our life!

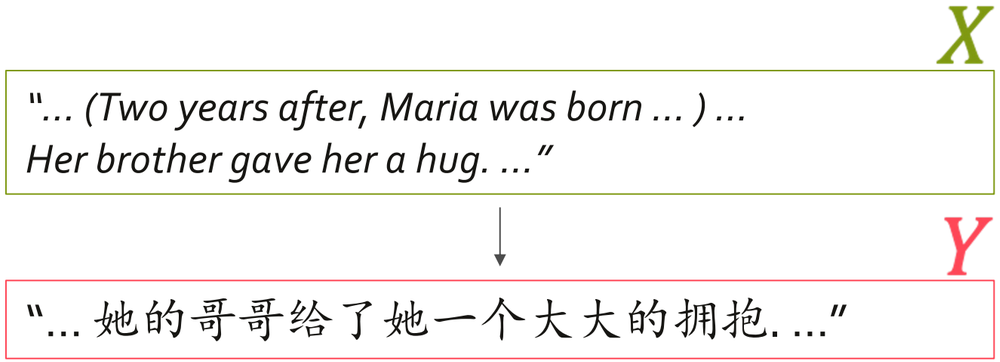
- $P(Y|X) = P(y_1, \cdots, y_m | x_1, \cdots, x_n) = \prod_{t=1}^{m} p(y_t | X, y_1, \cdots, y_{t-1})$, when Y is text and X is text.



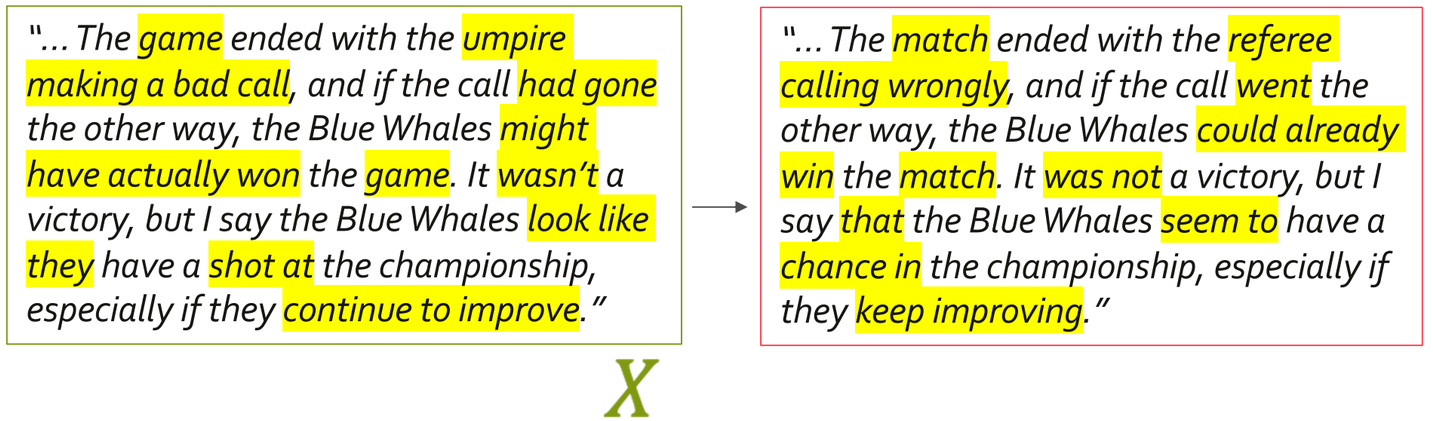
Dialog systems



Summarization

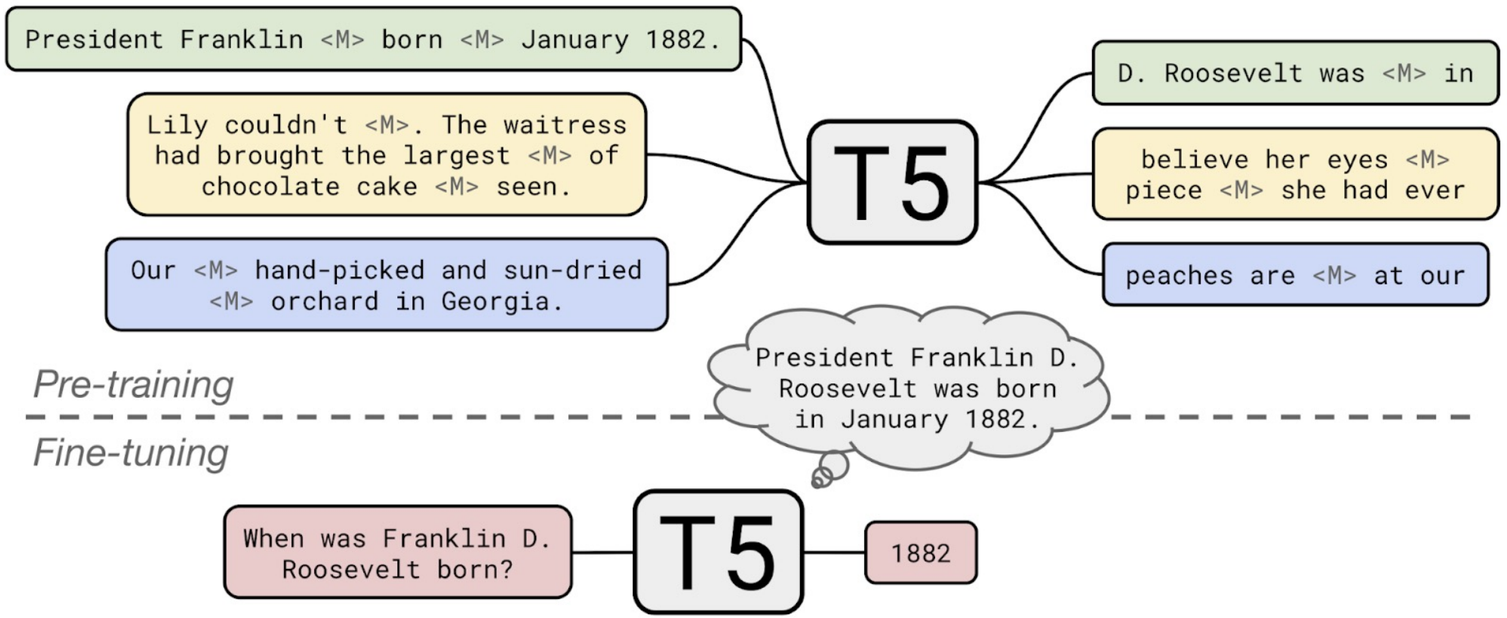


Machine Translation



Paraphrasing

Fine-tune PLMs on Downstream Tasks



Microsoft
UniLM, Turing-NLG

facebook
BART

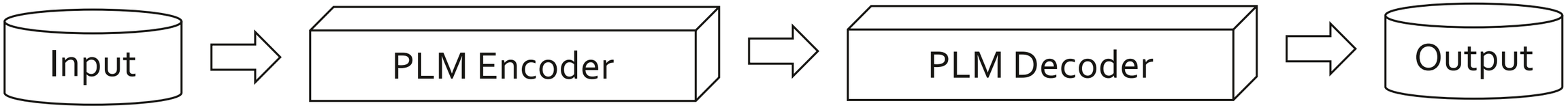
OpenAI
GPT-2, GPT-3

NVIDIA
Megatron LM

Google
T5

close-book

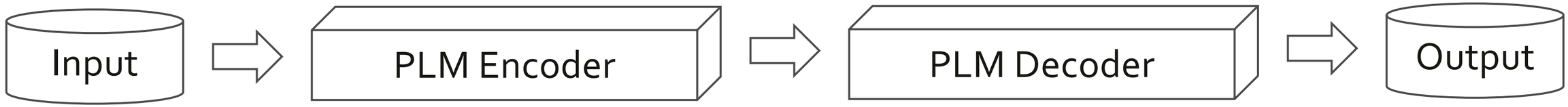
- Fine-tuning PLMs with *input-output* pairs of target data is the dominant paradigm in NLP research.



Input query: Miami Beach in Florida borders which ocean? Output Answer: Atlantic Ocean

Fine-tune PLMs on Downstream Tasks

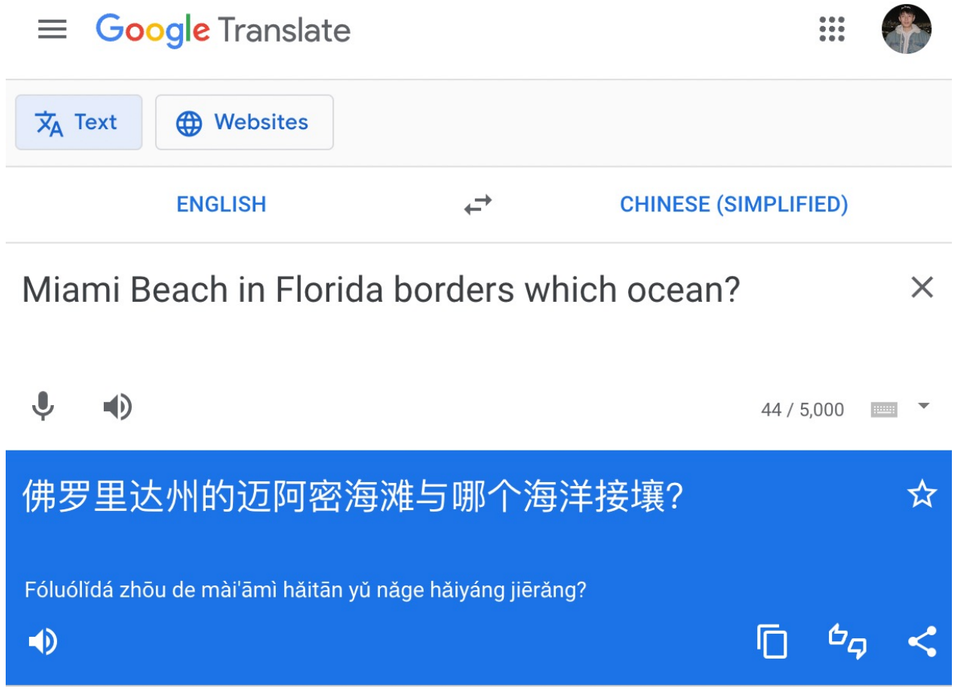
-- Fine-tuning PLMs with *input-output* of target data is the dominant paradigm.



- **Machine Translation**

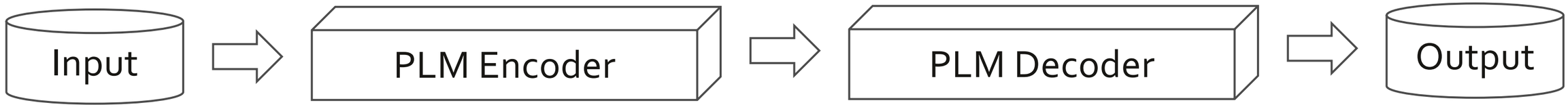
Input: Miami Beach in Florida borders which ocean?

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



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

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.

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

Why knowledge is needed in NLG?

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

-- Query: Who did Hawaii belong to before 1895?

Hawaiian Kingdom

Question-answer gap

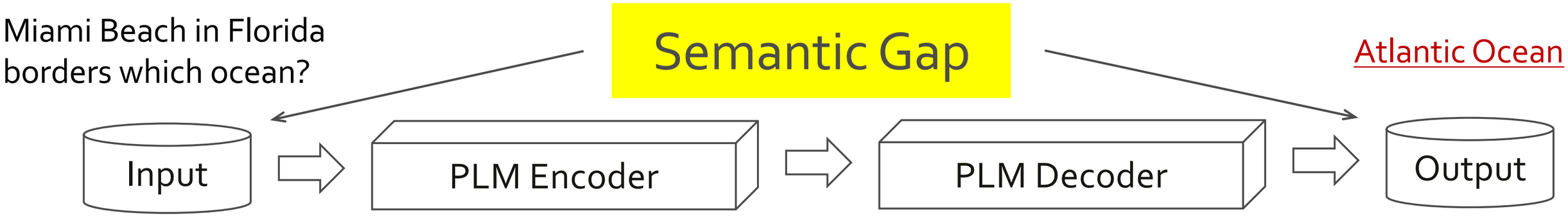
Hawaii is the most recent state … **On Jan 17, 1895**. The United States Minister to the **Hawaiian Kingdom** conspired with U.S. citizens to overthrow the monarchy.

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 close-book does not work?

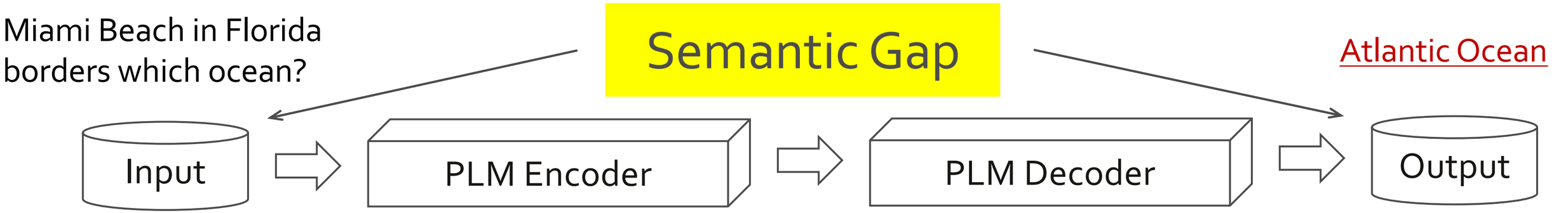


- When fine-tuned on the downstream tasks, the LMs might forget previously learned knowledge during pre-training, leading to catastrophic forgetting.

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

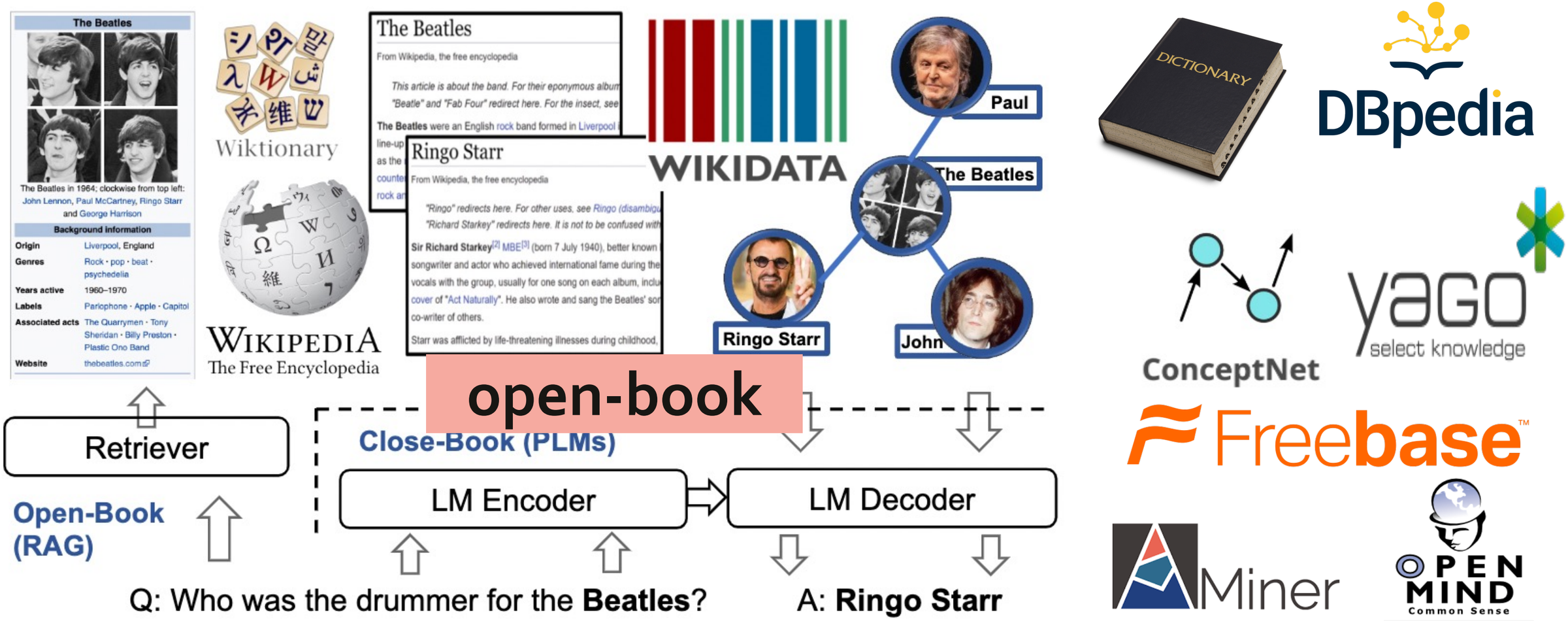
[1] Question and Answer Test-Train Overlap in Open-Domain QA Datasets. EACL 2021.

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, PaLM).

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)

Unstructured Knowledge
(i.e., grounded document)

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

Encyclopedic Knowledge
(i.e., Wikipedia, Aminer)

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

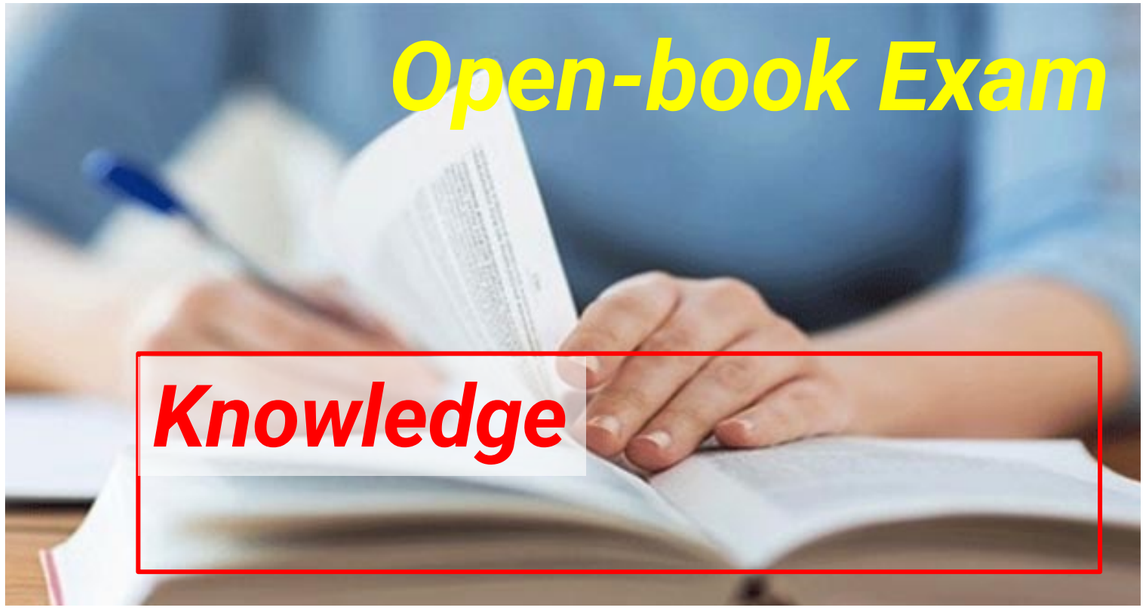
A Survey of Knowledge-Intensive Natural Language Processing
(Yin et al., on arXiv 2022)

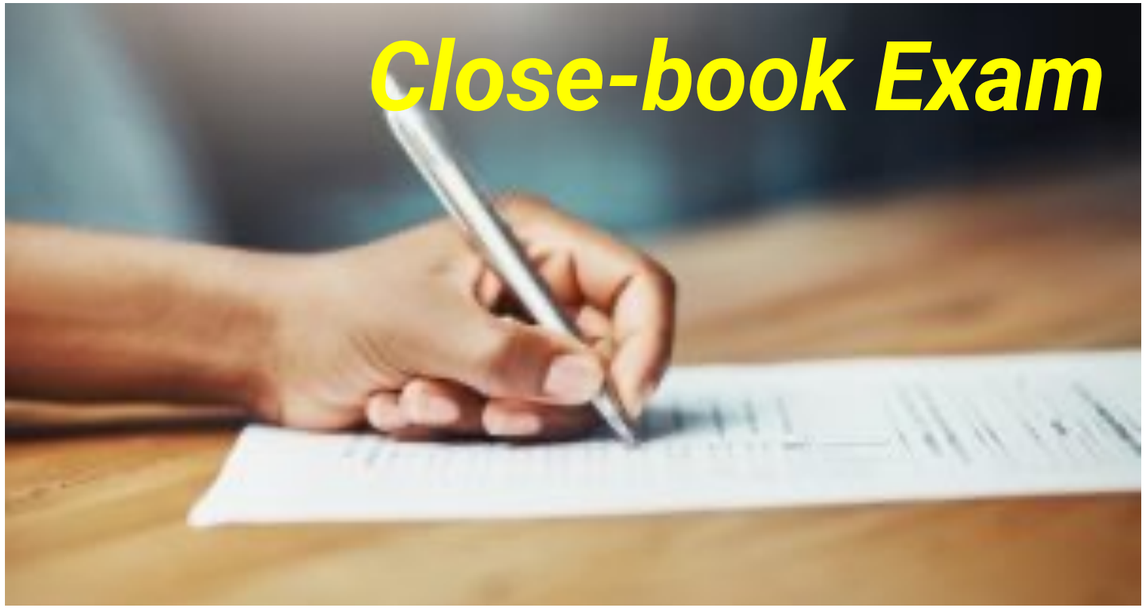
Summary: Close-book & Open-book

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-3 (e.g., ChatGPT).

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 RAG, FiD, KGFiD, RePLUG, (*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.

Papers list

-- Retrieval augmented NLG model:

[1] Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. Neurips 2020

[2] Leveraging passage retrieval with generative models for open domain question answering. EACL 2021

-- Retrieval augmented NLG model + LLM (e.g., GPT-3):

[3] Generate rather than Retrieve: Large Language Models are Strong Context Generators. ICLR 2023

[4] REPLUG: Retrieval-Augmented Black-Box Language Models. ArXiv 2023

-- Knowledge graph augmented NLG model:

[5] KG-FiD: Infusing Knowledge Graph in Fusion-in-Decoder for Open-Domain Question Answering. ACL 2022

[6] Grape: Knowledge Graph Enhanced Passage Reader for Open-domain Question Answering. EMNLP 2022

- -Memory augmented NLG model:

[7] A Unified Encoder-Decoder Framework with Entity Memory. EMNLP 2022

-- Knowledge augmented NLG model applications:

[8] Diversifying Content Generation with Mixture of Knowledge Graph Experts. ACL 2022

[9] Retrieval Augmentation for Commonsense Reasoning: A Unified Approach. EMNLP 2022

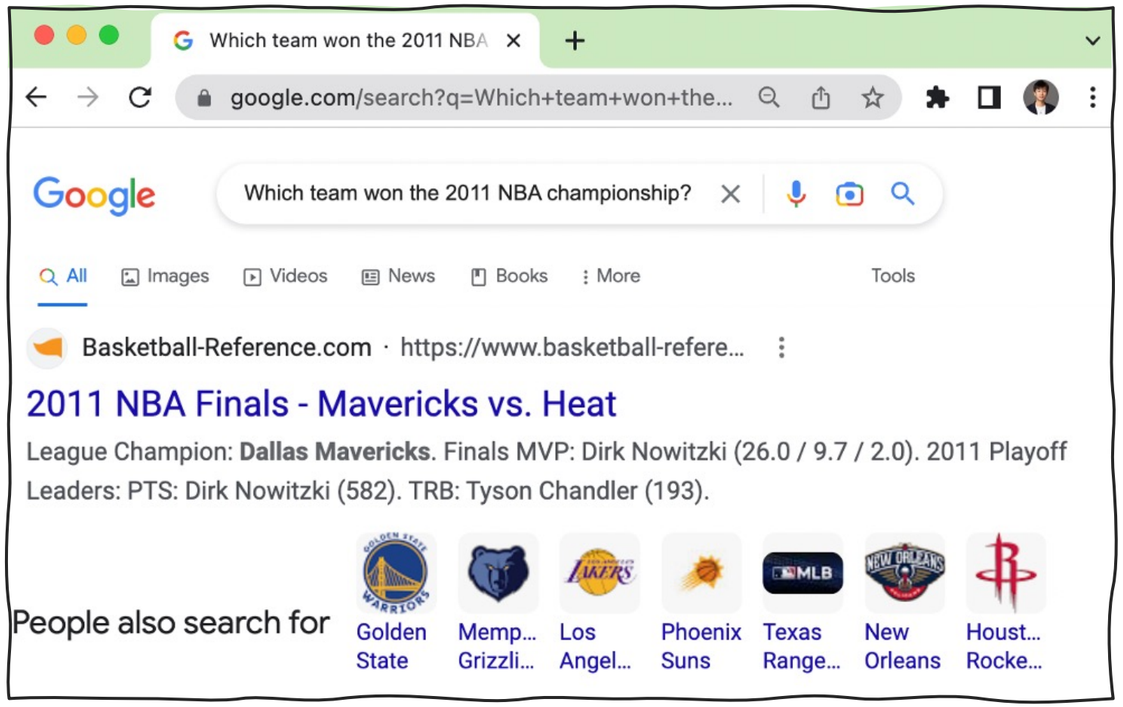
[10] Retrieval-Augmented Multimodal Language Modeling. ArXiv 2022

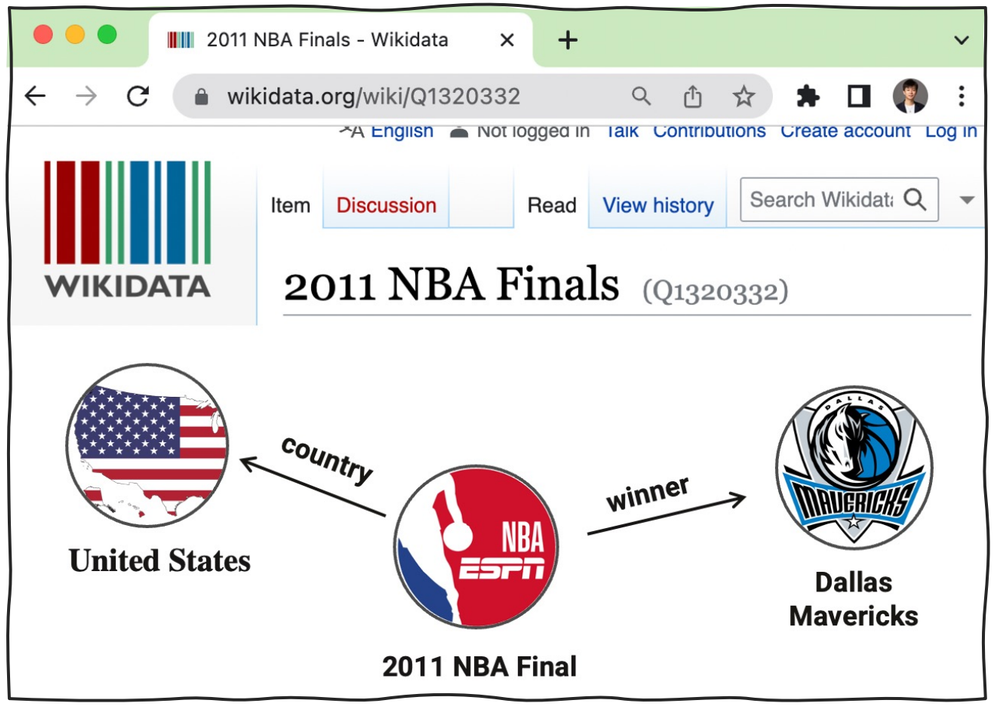
Research direction 1: Retrieval-augmented Language Models

Which team won the 2010-2011 NBA championship?

Dallas Mavericks

Input query → IR system → Retrieved information → Reader → Output answer

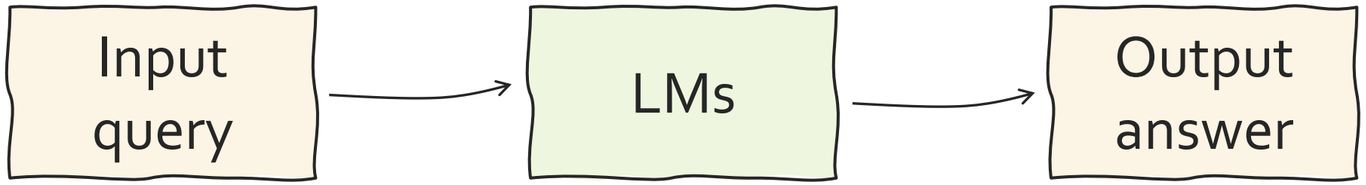




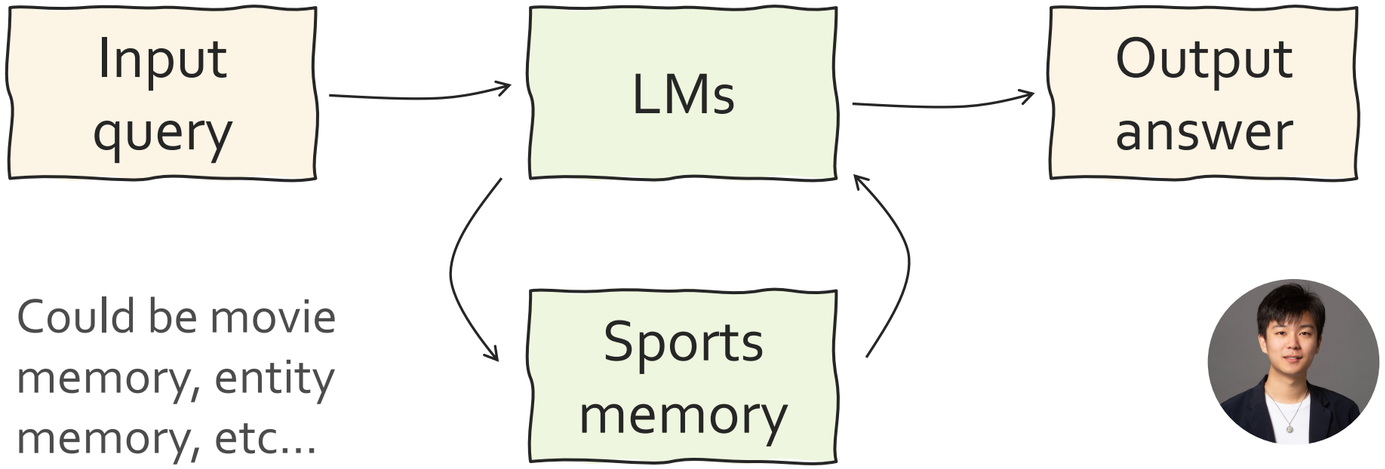
Research direction 2: Information seeking via GPT-3

Which team won the 2010-2011 NBA championship?

The Dallas Mavericks won the NBA championship in 2010-2011 season.



Research direction 3: Language Models + Memory



Could be movie memory, entity memory, etc…

Dallas Mavericks. I'm a Miami Heat fan, but they lost the finals. I am very frustrated.

Papers list

-- Retrieval augmented NLG model:

[1] Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. Neurips 2020

[2] Leveraging passage retrieval with generative models for open domain question answering. EACL 2021

-- Retrieval augmented NLG model + LLM (e.g., GPT-3):

[3] Generate rather than Retrieve: Large Language Models are Strong Context Generators. ICLR 2023

[4] REPLUG: Retrieval-Augmented Black-Box Language Models. ArXiv 2023

-- Knowledge graph augmented NLG model:

[5] KG-FiD: Infusing Knowledge Graph in Fusion-in-Decoder for Open-Domain Question Answering. ACL 2022

[6] Grape: Knowledge Graph Enhanced Passage Reader for Open-domain Question Answering. EMNLP 2022

- -Memory augmented NLG model:

[7] A Unified Encoder-Decoder Framework with Entity Memory. EMNLP 2022

-- Knowledge augmented NLG model applications:

[8] Diversifying Content Generation with Mixture of Knowledge Graph Experts. ACL 2022

[9] Retrieval Augmentation for Commonsense Reasoning: A Unified Approach. EMNLP 2022

[10] Retrieval-Augmented Multimodal Language Modeling. ArXiv 2022

Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks

The 2020 Conference on Neural Information Processing Systems

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni,
Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis,
Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, Douwe Kiela

RAG: Retrieval-augmented Generation

(1) It is the first retrieval-augmented generation work. The model is named RAG.

(2) RAG combines a pre-trained retriever (Query Encoder + Document Index) with a pre-trained seq2seq model (Generator) and fine-tune end-to-end.

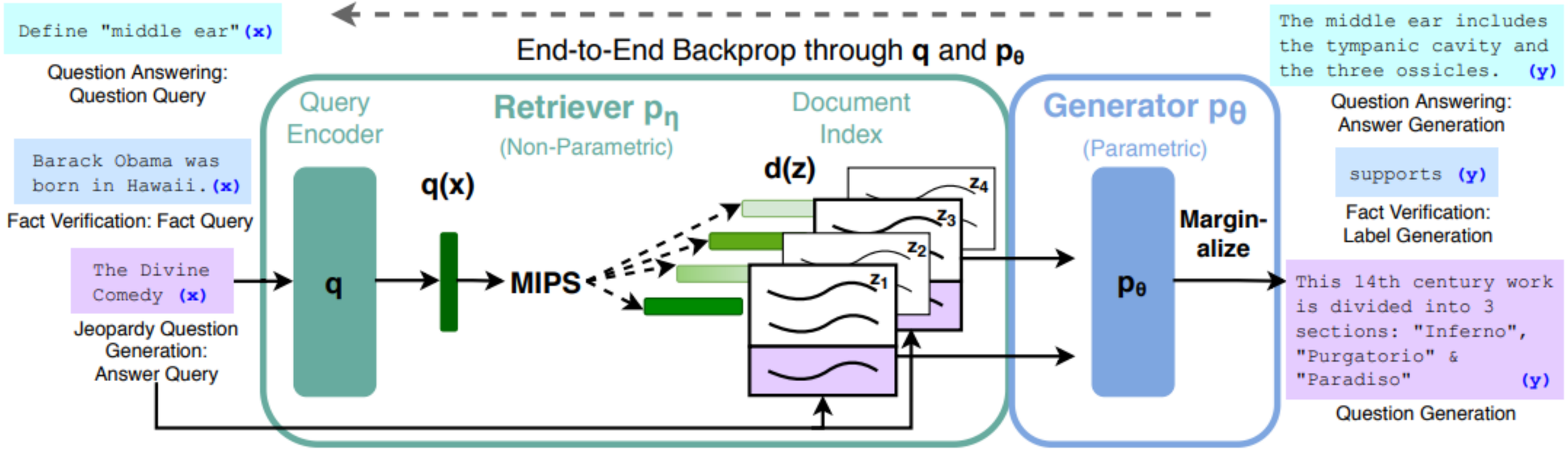


Figure: Overview of retrieval-augmented generation (RAG) approach. Figure is copied from paper figure 1.

[1] Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks, Neruips'20

RAG: Model frameworks

Model variant 1 -- RAG-token: top K documents are retrieved using the retriever, and the generator produces the output sequence probability for each document, then marginalized.

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

Model variant 2 -- RAG-Sequence: top K documents are retrieved using the retriever, and then the generator produces a distribution for the next output token for each document, before marginalizing, and repeating the process with the following output token.

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

[1] Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks, Neruips'20

RAG: Downstream task performance

Dataset: TriviaQA, 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]	34.5	- /50.1	37.4	-
	T5-11B+SSM[52]	36.6	- /60.5	44.7	-
Open Book	REALM [20]	40.4	- / -	40.7	46.8
	DPR [26]	41.5	**57.9**/ -	41.1	50.6
	RAG-Token	44.1	55.2/66.1	**45.5**	50.0
	RAG-Seq.	**44.5**	56.8/**68.0**	45.2	**52.2**

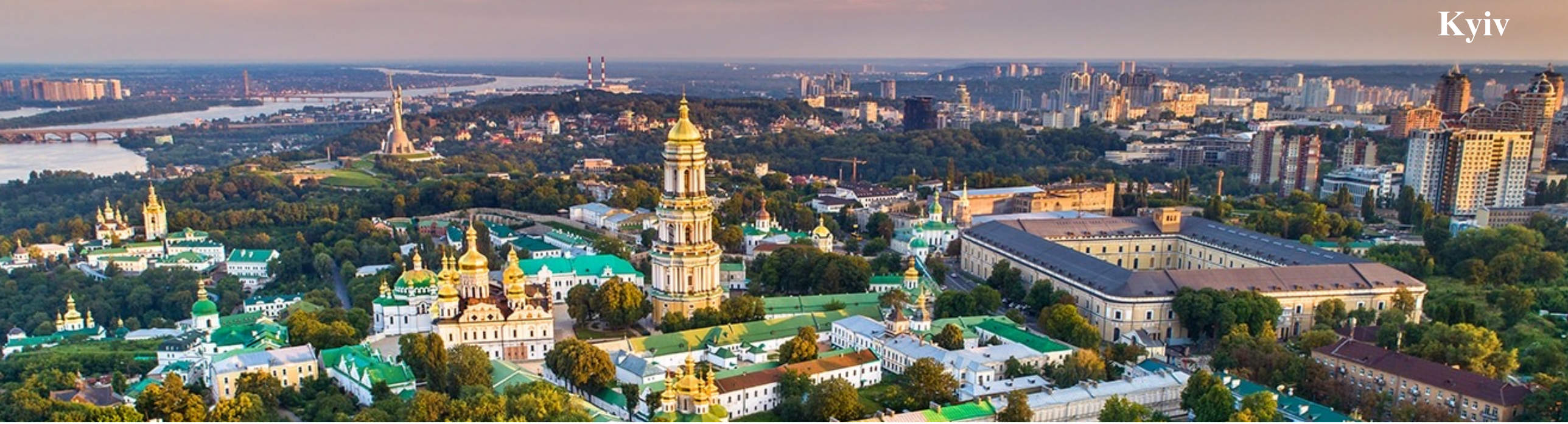
Table 2: Generation and classification Test Scores. MS-MARCO SotA is [4], FEVER-3 is [68] and FEVER-2 is [57] *Uses gold context/evidence. Best model without gold access underlined.

Model	Jeopardy B-1	Jeopardy QB-1	MSMARCO R-L	MSMARCO B-1	FVR3 Label Acc.	FVR2 Label Acc.
SotA	-	-	**49.8***	**49.9***	**76.8**	**92.2***
BART	15.1	19.7	38.2	41.6	64.0	81.1
RAG-Tok.	**17.3**	**22.2**	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 TriviaQA

RAG > Close-book (BART)

[1] Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks, Neruips'20



Leveraging Passage Retrieval with Generative Models for Open-domain Question Answering

The 2021 European Chapter of the Association for Computational Linguistics

Gautier Izacard, Edouard Grave

FiD: Fusion-in-decoder

Why not RAG: RAG suffers from the input sequence length limitation (max:1024) and high computation cost (quadratic to the sequence length).

How FiD overcome the above drawback: FiD processed passages independently in the encoder, performed attention over all the retrieved passages

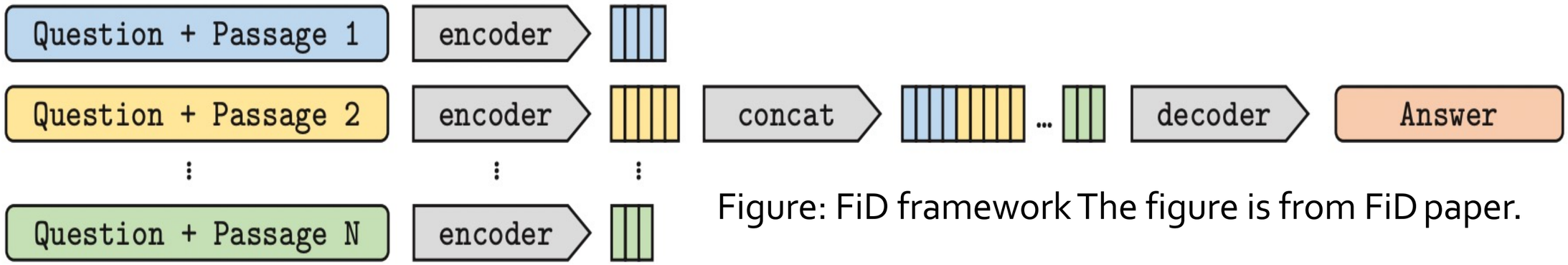


Figure: FiD framework The figure is from FiD paper.

[2] Leveraging Passage Retrieval with Generative Models for Open-domain QA. In EACL'21

FiD: Downstream QA performance

Model	NQ	TriviaQA		SQuAD Open	
	EM	EM	EM	EM	F1
DrQA (Chen et al., 2017)	-	-	-	29.8	-
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)	**51.4**	**67.6**	**80.1**	**56.7**	63.2

+10% on average

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

FiD is by far the most popular retrieval augmented model, owing to its simple architecture and strong performance.

[2] Leveraging Passage Retrieval with Generative Models for Open-domain QA. In EACL'21

Papers list

-- Retrieval augmented NLG model:

[1] Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. Neurips 2020

[2] Leveraging passage retrieval with generative models for open domain question answering. EACL 2021

-- Retrieval augmented NLG model + LLM (e.g., GPT-3):

[3] Generate rather than Retrieve: Large Language Models are Strong Context Generators. ICLR 2023

[4] REPLUG: Retrieval-Augmented Black-Box Language Models. ArXiv 2023

-- Knowledge graph augmented NLG model:

[5] KG-FiD: Infusing Knowledge Graph in Fusion-in-Decoder for Open-Domain Question Answering. ACL 2022

[6] Grape: Knowledge Graph Enhanced Passage Reader for Open-domain Question Answering. EMNLP 2022

- -Memory augmented NLG model:

[7] A Unified Encoder-Decoder Framework with Entity Memory. EMNLP 2022

-- Knowledge augmented NLG model applications:

[8] Diversifying Content Generation with Mixture of Knowledge Graph Experts. ACL 2022

[9] Retrieval Augmentation for Commonsense Reasoning: A Unified Approach. EMNLP 2022

[10] Retrieval-Augmented Multimodal Language Modeling. ArXiv 2022

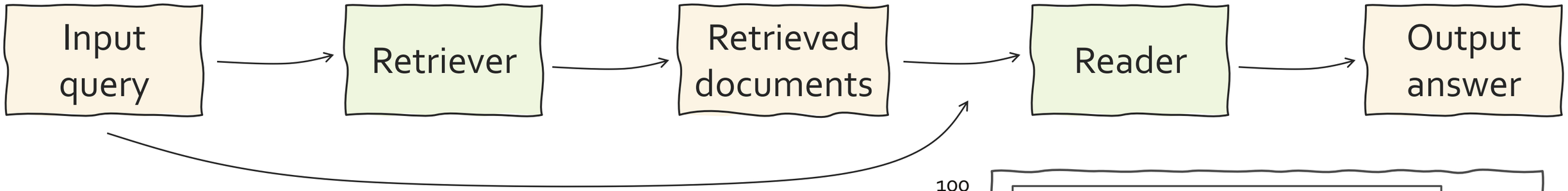
Generate rather than Retrieve: Large Language Models are Strong Context Generators

The 2023 International Conference on Learning Representations

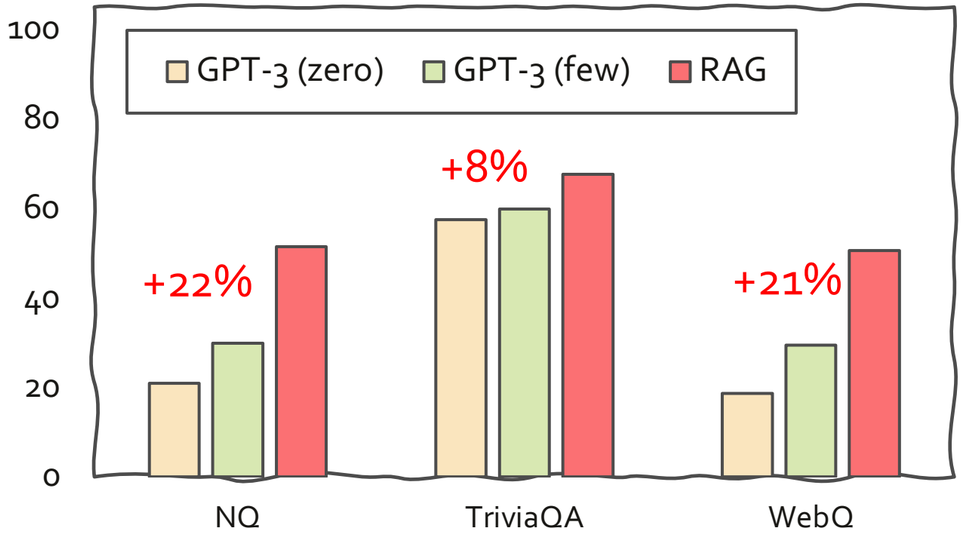
Wenhao Yu, Dan Iter, Shuohang Wang, Yichong Xu, Mingxuan Ju, Soumya Sanyal, Chenguang Zhu, Michael Zeng, Meng Jiang

GenRead: Generate rather than Retriever

Recall: how retrieval-augmented LMs work? – short as RAG.



Recall: how GPT-3 work?



No! GPT-3 is still left behind by RAG models! But why?

[1] Generate rather than Retrieve: Large Language Models are Strong Context Generators. ICLR 2023

GenRead: Do GPT-3 know the answer?

Who got the first Nobel prize in physics?

Albert Einstein ✗

Please provide a background document for the following question.
Who got the first Nobel prize in physics?

The first Nobel Prize in Physics was awarded in 1901 to **Wilhelm Conrad Rontgen** in recognition of the extraordinary services he has rendered by the discovery of the remarkable rays subsequently named after him. -- by text-davinci-002 ✓

Yes. The answer is contained in the generated document!

[1] Generate rather than Retrieve: Large Language Models are Strong Context Generators. ICLR 2023

GenRead: Why generate document, not answer?

-- During the pre-training of large language models (e.g., GPT-3), document-level text is the dominant type of input and output. Therefore,

(1) the goal of generating document-level context for a question can serve as an intermediate step that simulate the language modeling pre-training.

(2) the task of directly generating answers prevents models from effectively recalling knowledge from their memory.

We humans do the same!

What is the tenth decimal of π?

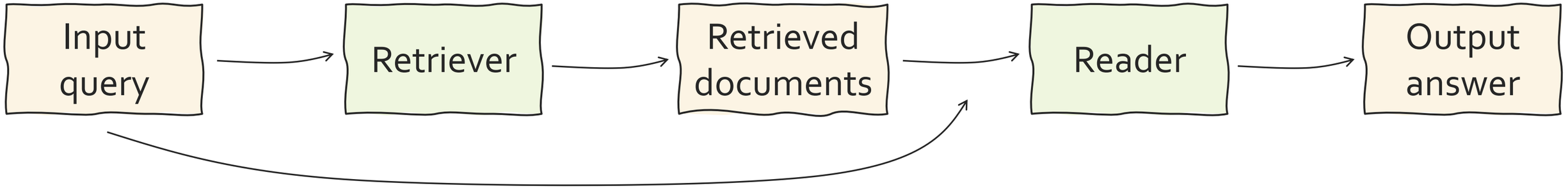
π is 3.14 15926 535. It is **5!** [2]

[1] Generate rather than Retrieve: Large Language Models are Strong Context Generators. ICLR 2023

[2] Recitation-Augmented Language Models. ICLR 2023

GenRead: Generate-then-Read pipeline!

Recall: how retrieval-augmented LMs work?



How Generate-then-Read (GenRead) work?

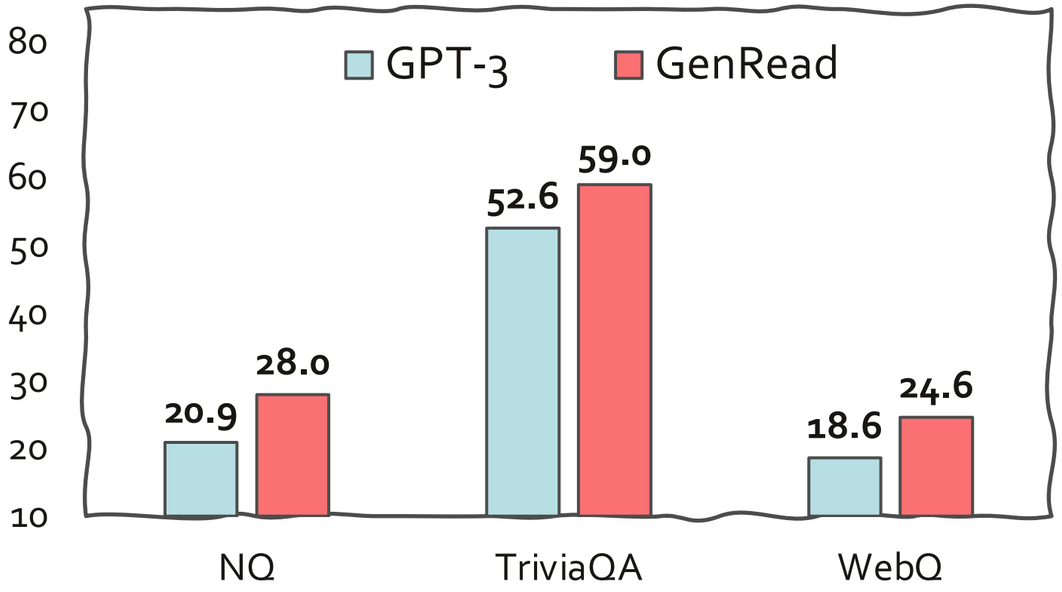
Replacing a retriever with GPT-3 generator!

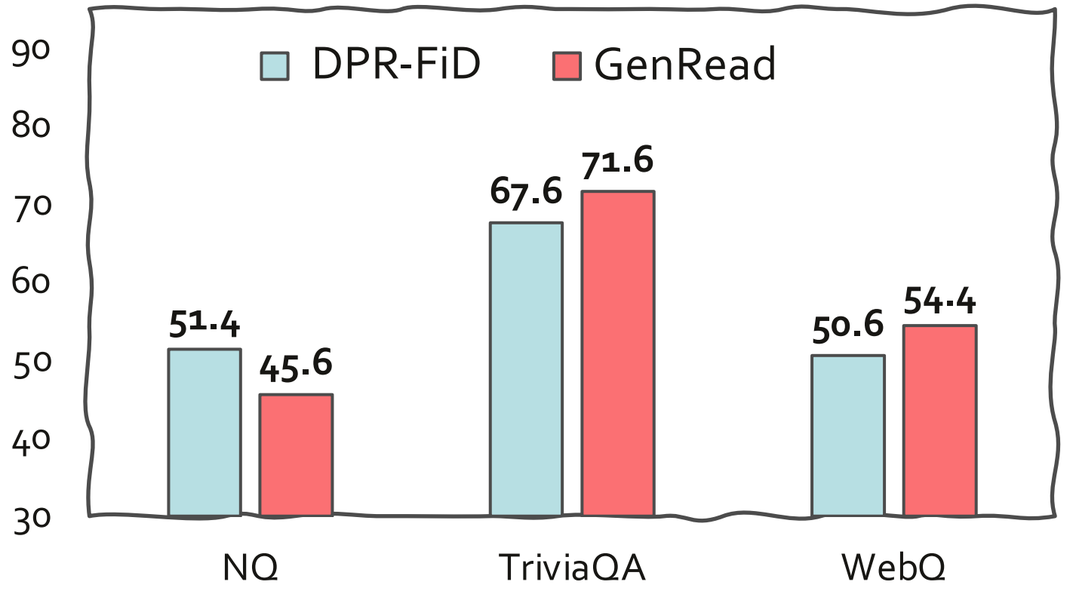
[1] Generate rather than Retrieve: Large Language Models are Strong Context Generators. ICLR 2023

GenRead: Downstream QA performance

(w/o training set)	Retriever	Reader
Baseline (GPT)	✕	InstructGPT
Ours (GenRead)	InstrcutGPT	InstructGPT

(with training set)	Retriever	Reader
Baseline (DPR-FiD)	DPR	FiD
Ours (GenRead)	InstrcutGPT	FiD





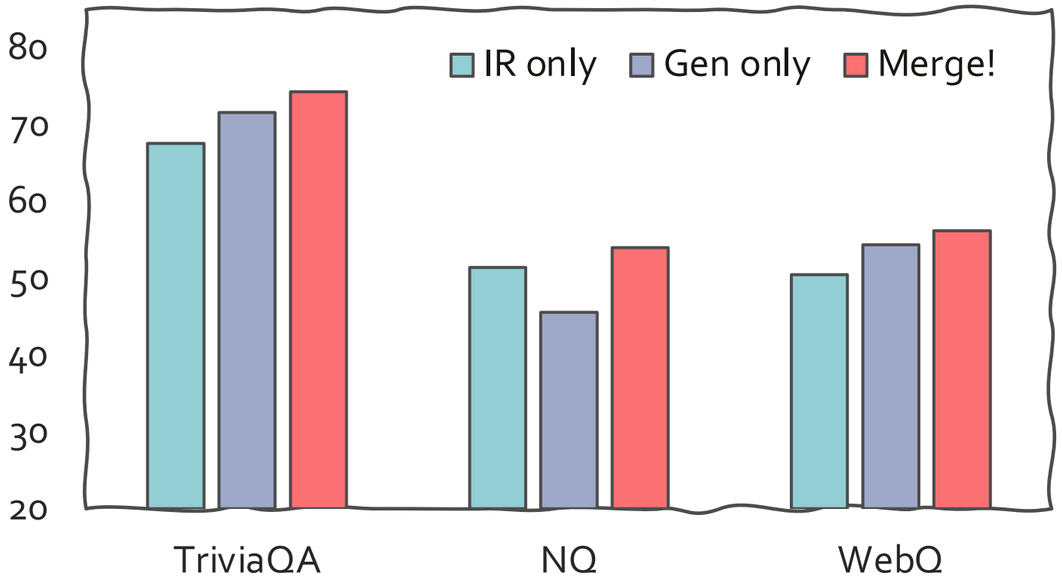
Zero-shot setting: +6.5% over InstructGPT!

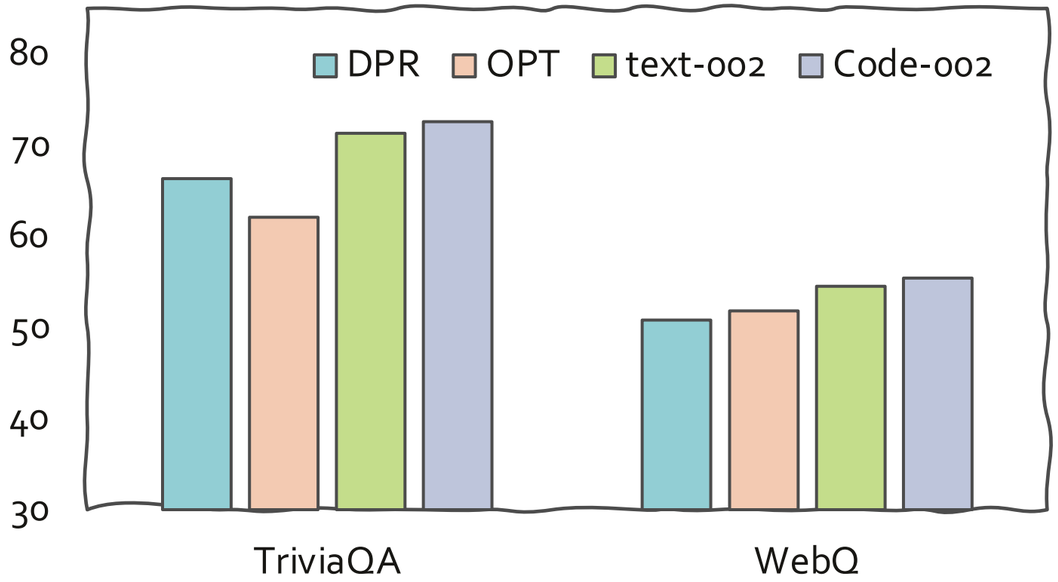
Supervised setting: +0.7% over SoTA!

Generate-then-read beats SoTA under both settings!

[1] Generate rather than Retrieve: Large Language Models are Strong Context Generators. ICLR 2023

GenRead: Two observations





Merge IR + Gen > IR only and Gen only

IR only: **56.5** Gen only: **57.1** Merge: **61.5**

GenRead works well on different LLMs

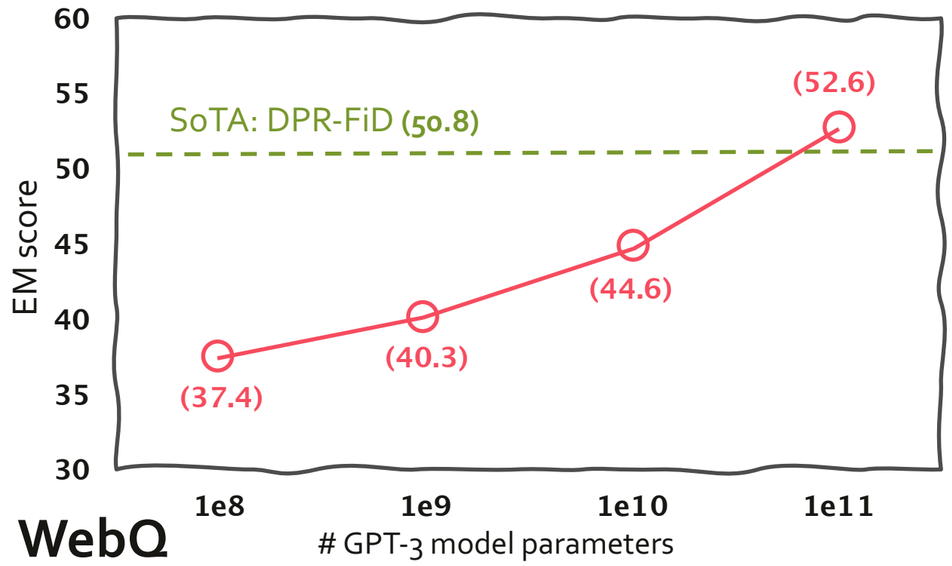
Code-davinci-002 (OpenAI) works best!

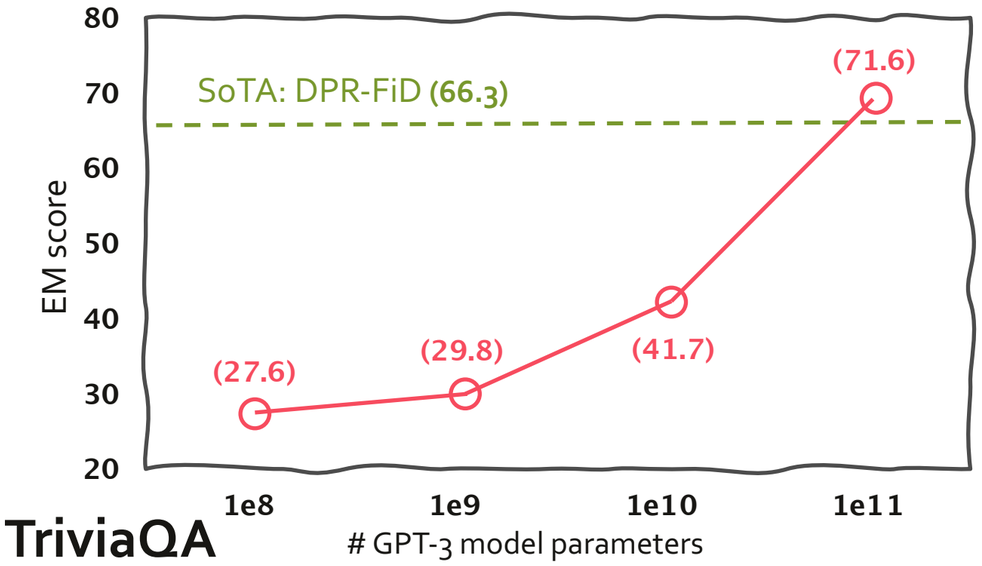
Combining retrieved and generated documents works better

Generate-then-read is a general pipeline for many LLMs

[1] Generate rather than Retrieve: Large Language Models are Strong Context Generators. ICLR 2023

GenRead: Scaling with number of GPT-3 parameters from 100 million to 175 billion





The performance on both datasets continues to improve as the GPT-3 model size increase, as does the slope.

Only with the largest size GPT-3, GenRead can outperform the DPR-FiD. So, the ability only presents in LLMs [1].

[1] Generate rather than Retrieve: Large Language Models are Strong Context Generators. ICLR 2023

Retrieval-augmented LMs vs. Very Large LMs

Pros. of retrieval-augmented LMs

-- Document corpus is updatable. New information can be easily plugged in.

-- More efficient! Usually, a retriever model is small, not requiring to store all world knowledge into model parameters.

Pros. of very large LMs

-- Better knowledge storage (>100B), strong deductive reasoning capabilities.

-- Generated information are more specific to the given input, i.e., customized output.

Cons. of retrieval-augmented LMs

-- Documents are fixed. They might contain noisy or irrelevant information.

-- Retrievers are based on $simi(\vec{q}, \vec{p})$:

- Shallow interactions of $(\vec{q}, \vec{p})$
- Different semantic space of $\vec{q}, \vec{p}$

Cons. of very large LMs

-- (Generative) Large language models still often hallucinate wrong information.

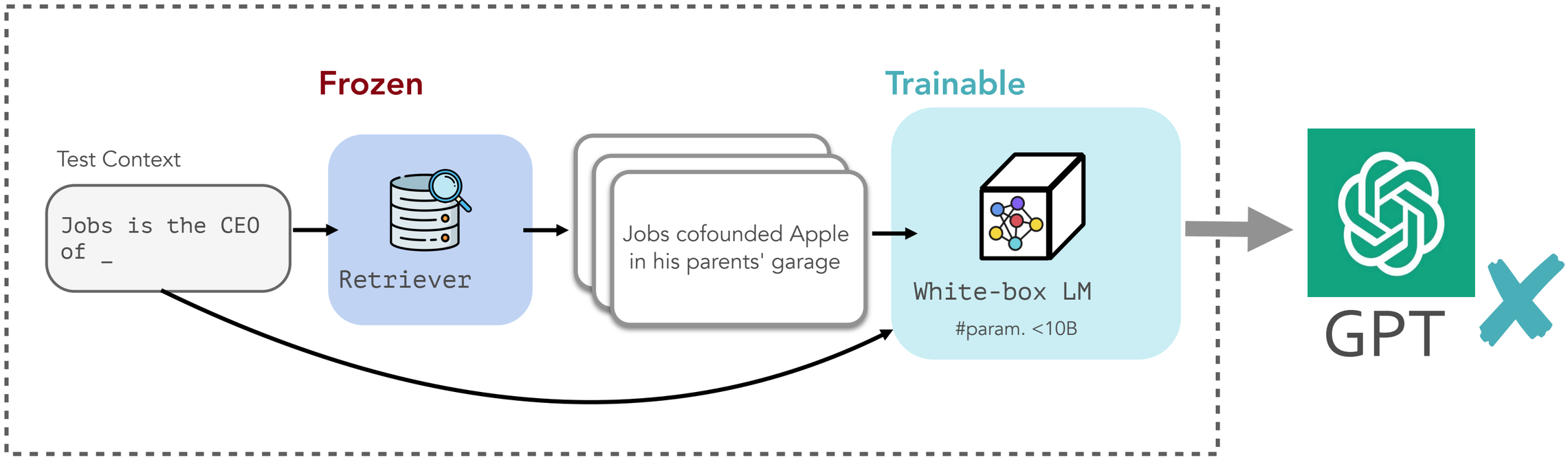
-- Updating new information into model parameters is hard, which if often non-affordable in academic labs.

RePLUG: Retrieval-augmented Black-Box LM

Weijia Shi, Sewon Min, Michihiro Yasunaga, Minjoon Seo, Rich James, Mike Lewis, Luke Zettlemoyer, Wen-tau Yih

RePLUG: Retrieval-augmented Black-Box LM

Previous Framework: RETRO (Borgeaud, et al. 2022), Atlas (Izacard, et al. 2022)

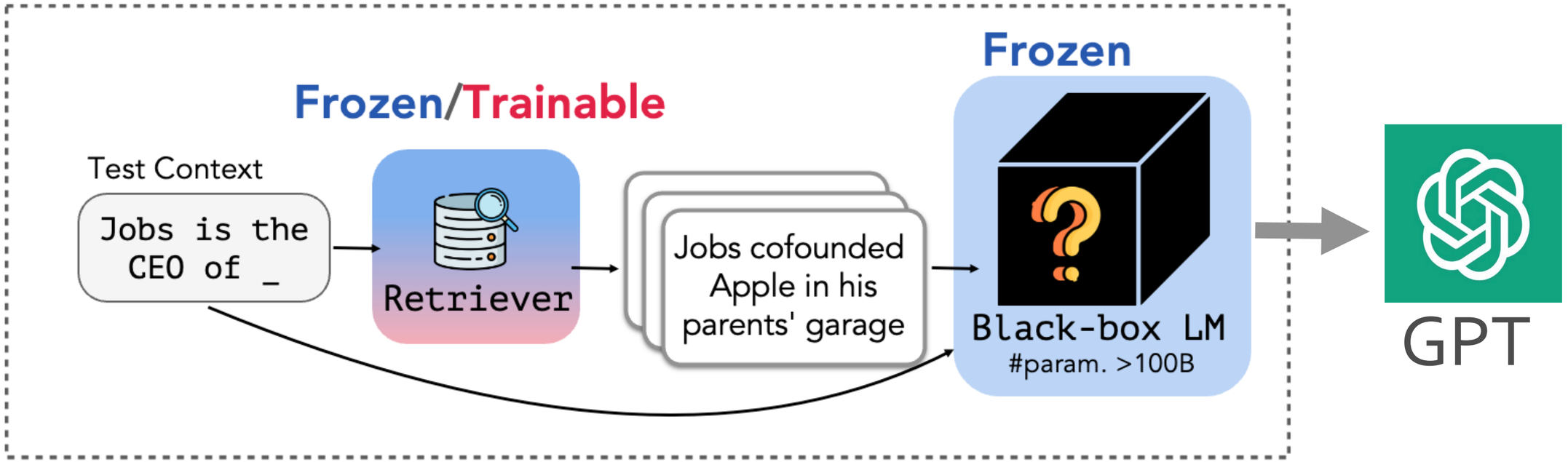


Black-box LLM is only accessible by API, and expensive to finetune

[1] REPLUG: Retrieval-Augmented Black-Box Language Models. ArXiv 2023

RePLUG: Retrieval-augmented Black-Box LM

RePLUG: enhancing large black-box language models with retrieval

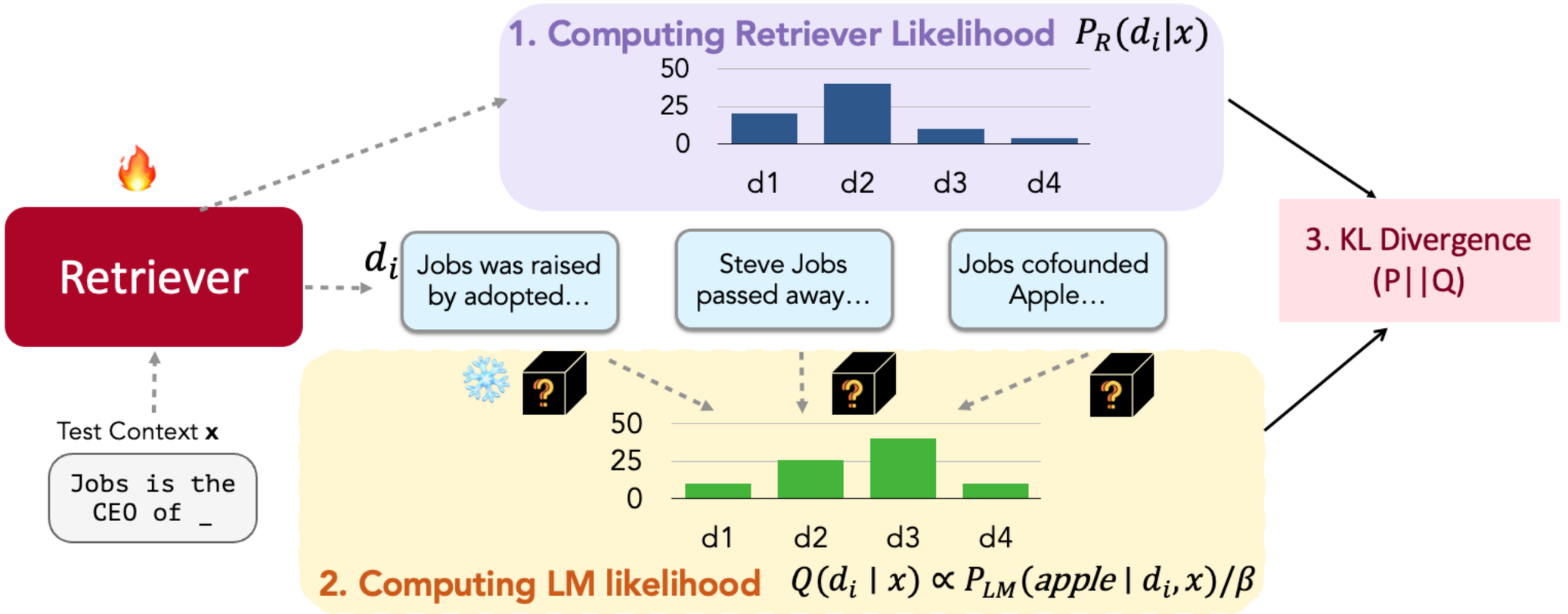


Taking advantages of both retrieval-augmented LMs and LLMs

[1] REPLUG: Retrieval-Augmented Black-Box Language Models. ArXiv 2023

RePLUG: Training the dense retriever

-- Step 3: Optimize KL divergence

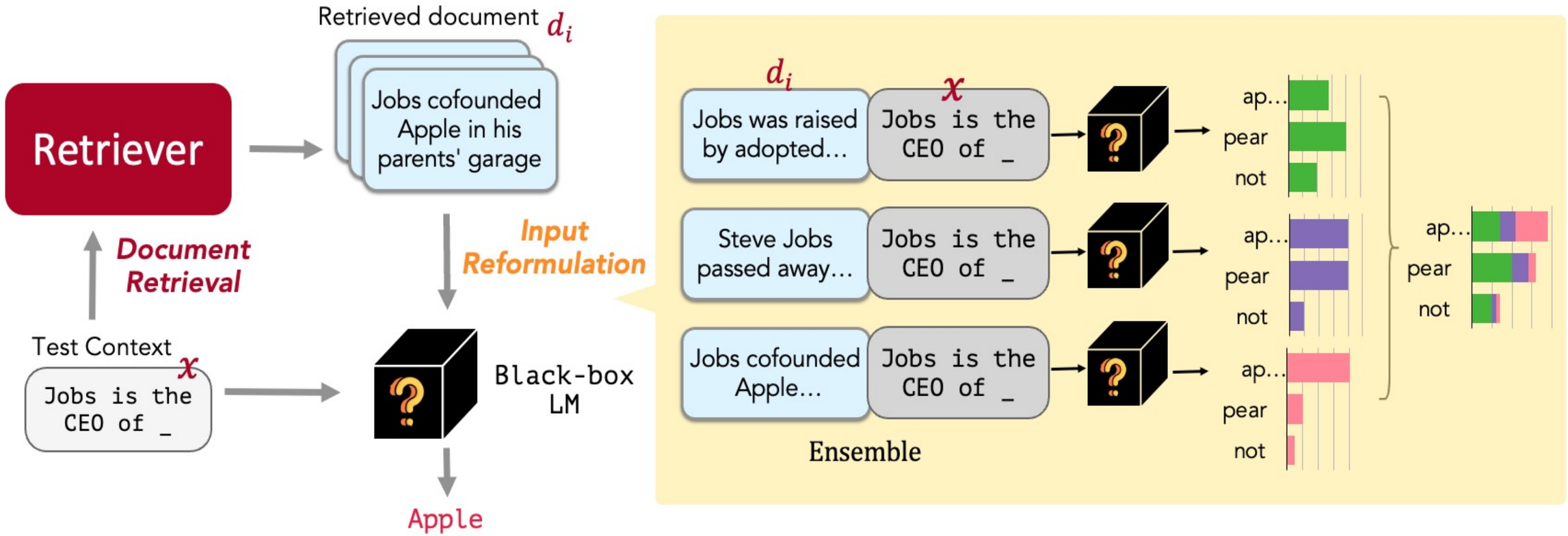


[1] REPLUG: Retrieval-Augmented Black-Box Language Models. ArXiv 2023

RePLUG: Inference

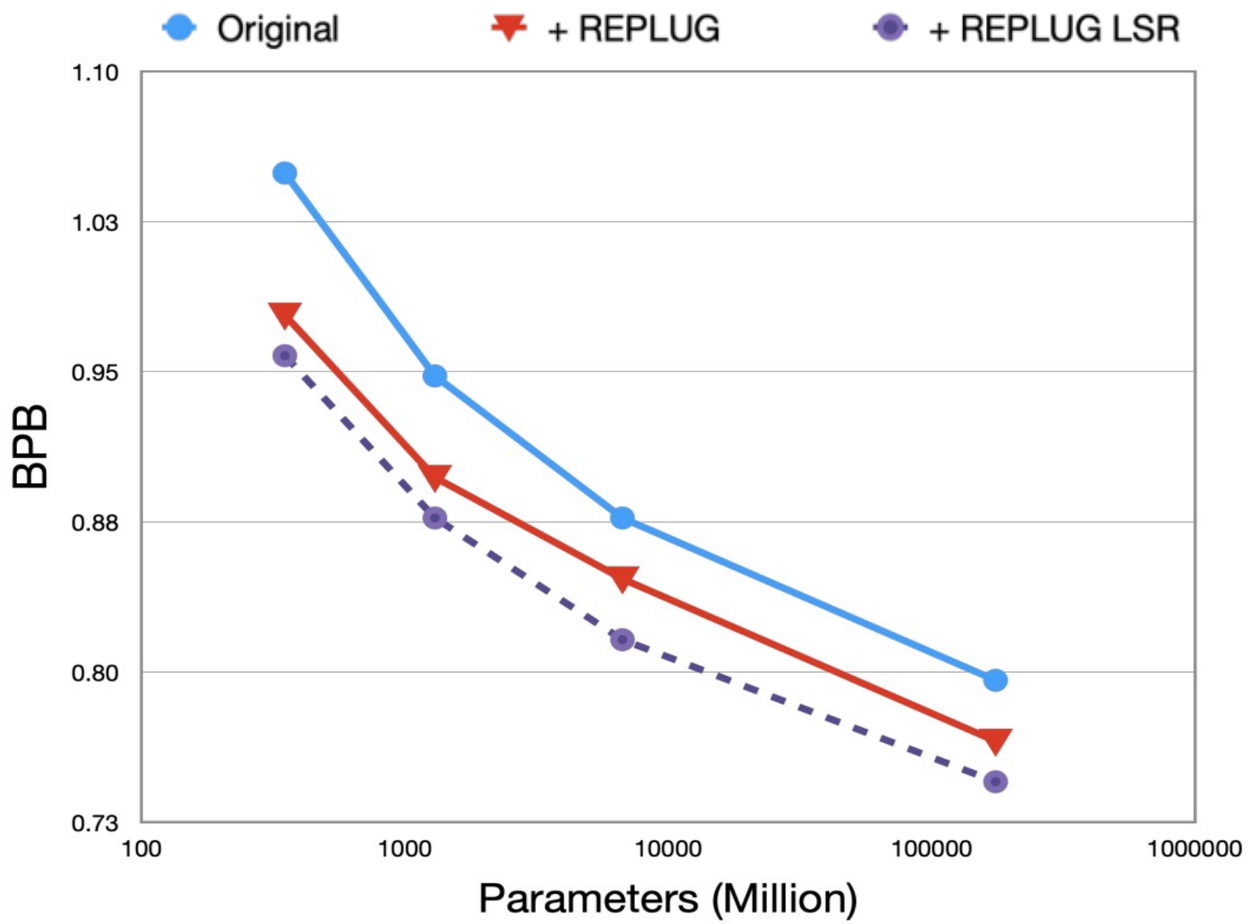
-- Step 1: Document Retrieval

-- Step 2: Information Fusion



[1] REPLUG: Retrieval-Augmented Black-Box Language Models. ArXiv 2023

RePLUG: Retrieval-augmented Black-Box LM



- Experiments on MMLU:

Models	Social.	STEM	All
Atlas	54.6	38.8	47.9
Codex	76.9	57.8	68.3
RePLUG	**79.9**	**58.9**	**71.8**

- Experiments on open-domain QA:

Models	NQ		TriviaQA	
	64-shot	Full	64-shot	Full
Atlas	42.4	60.4	73.6	79.8
Codex	40.6	-	74.5	-
RePLUG	**45.5**	**-**	**77.3**	**-**

RePLUG consistently enhanced the performance on different tasks

[1] REPLUG: Retrieval-Augmented Black-Box Language Models. ArXiv 2023

Conclusion and Future Works (1/3)

Are retrieval-augmented LMs still needed in the GPT-3 Era?

Yes, despite advancements made by GPT-3, Retrieval-augmented LMs still play an important role. They provide the ability to access updated information in real-time, then serving as input to the LLM, allowing the benefits of both IR and LLM to be leveraged.

Can retrieval documents be co-mbined with GPT-3 documents?

Yes. Combining retrieved and generated documents can further improve performance.

Conclusion and Future Works (2/3)

Can you give some future directions regarding to the topic of retrieval-augmented LMs and very large language models?

Future directions include combining the strengths of both generative and retrieval-based models, and the exploration of how these models can be applied to a wider range of NLP tasks and domains. Besides, there is ongoing research into developing more interpretable models to better understand how large language models make predictions, and how to ensure that they are making fair and unbiased decisions.

Conclusion and Future Works (3/3)

Can you give some future directions regarding to the topic of retrieval-augmented LMs and very large language models?

(1) **When trust retriever, and when trust GPT-3?** In [Mallen 2022] they found GPT-3 performs poorly on rare entities, significantly left behind by retrieval-augmented LMs.

(2) **Can we trust GPT-3 outputs?** incorporating a retriever that makes a judgment when the model is unsure of the answer.

(3) **Structured facts as condition?** incorporating structured facts as constraints to prompt GPT-3 generate faithful outputs.

[1] Mallen et al. When Not to Trust Language Models: Investigating Effectiveness and Limitations of Parametric and Non-Parametric Memories. arXiv 2022. [2] Under review.

Papers list

-- Retrieval augmented NLG model:

[1] Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. Neurips 2020

[2] Leveraging passage retrieval with generative models for open domain question answering. EACL 2021

-- Retrieval augmented NLG model + LLM (e.g., GPT-3):

[3] Generate rather than Retrieve: Large Language Models are Strong Context Generators. ICLR 2023

[4] REPLUG: Retrieval-Augmented Black-Box Language Models. ArXiv 2023

-- Knowledge graph augmented NLG model:

[5] KG-FiD: Infusing Knowledge Graph in Fusion-in-Decoder for Open-Domain Question Answering. ACL 2022

[6] Grape: Knowledge Graph Enhanced Passage Reader for Open-domain Question Answering. EMNLP 2022

- -Memory augmented NLG model:

[7] A Unified Encoder-Decoder Framework with Entity Memory. EMNLP 2022

-- Knowledge augmented NLG model applications:

[8] Diversifying Content Generation with Mixture of Knowledge Graph Experts. ACL 2022

[9] Retrieval Augmentation for Commonsense Reasoning: A Unified Approach. EMNLP 2022

[10] Retrieval-Augmented Multimodal Language Modeling. ArXiv 2022

KG-FiD: Infusing Knowledge Graph in Fusion-in-Decoder for Open-Domain Question Answering

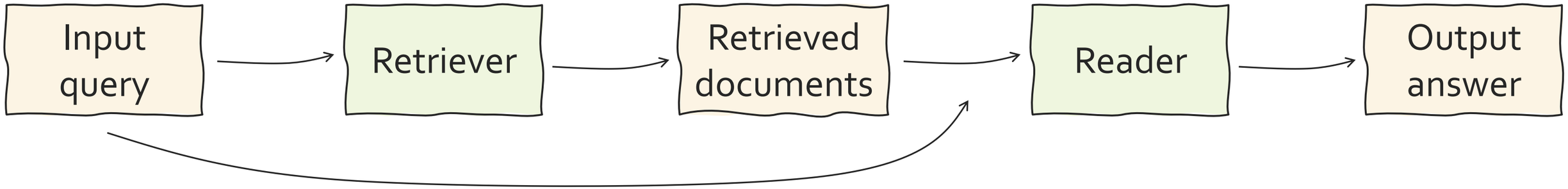
The 2022 Annual Meeting of the Association for Computational Linguistics

Donghan Yu, Chenguang Zhu, Yuwei Fang, Wenhao Yu, Shuohang Wang,
Yichong Xu, Xiang Ren, Yiming Yang, Michael Zeng

KG-FiD: Knowledge graph enhanced FiD

-- The independent assumption among passages in FiD is not justified. Notice that both the DPR retriever and the generative reader of FiD perform independent encoding of the retrieved passages, which means that they cannot leverage the semantic relationship among passages for passage embedding and answer generation even if such relational knowledge is available.

Recall: how FiD work?

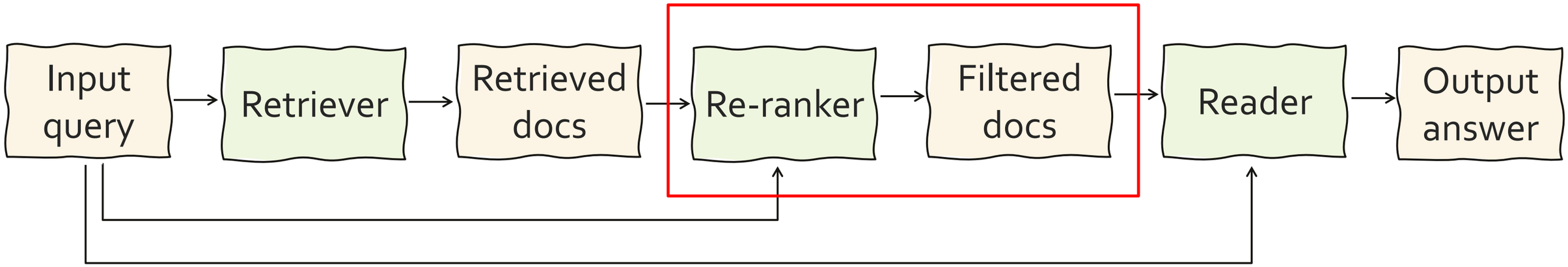


[1] KG-FiD: Infusing Knowledge Graph in Fusion-in-Decoder for Open-Domain Question Answering. ACL 2022

KG-FiD: Knowledge graph enhanced FiD

-- The independent assumption among passages in FiD is not justified. Notice that both the DPR retriever and the generative reader of FiD perform independent encoding of the retrieved passages, which means that they cannot leverage the semantic relationship among passages for passage embedding and answer generation even if such relational knowledge is available.

How **KG-FiD** work?



[1] KG-FiD: Infusing Knowledge Graph in Fusion-in-Decoder for Open-Domain Question Answering. ACL 2022

KG-FiD: Knowledge graph enhanced FiD

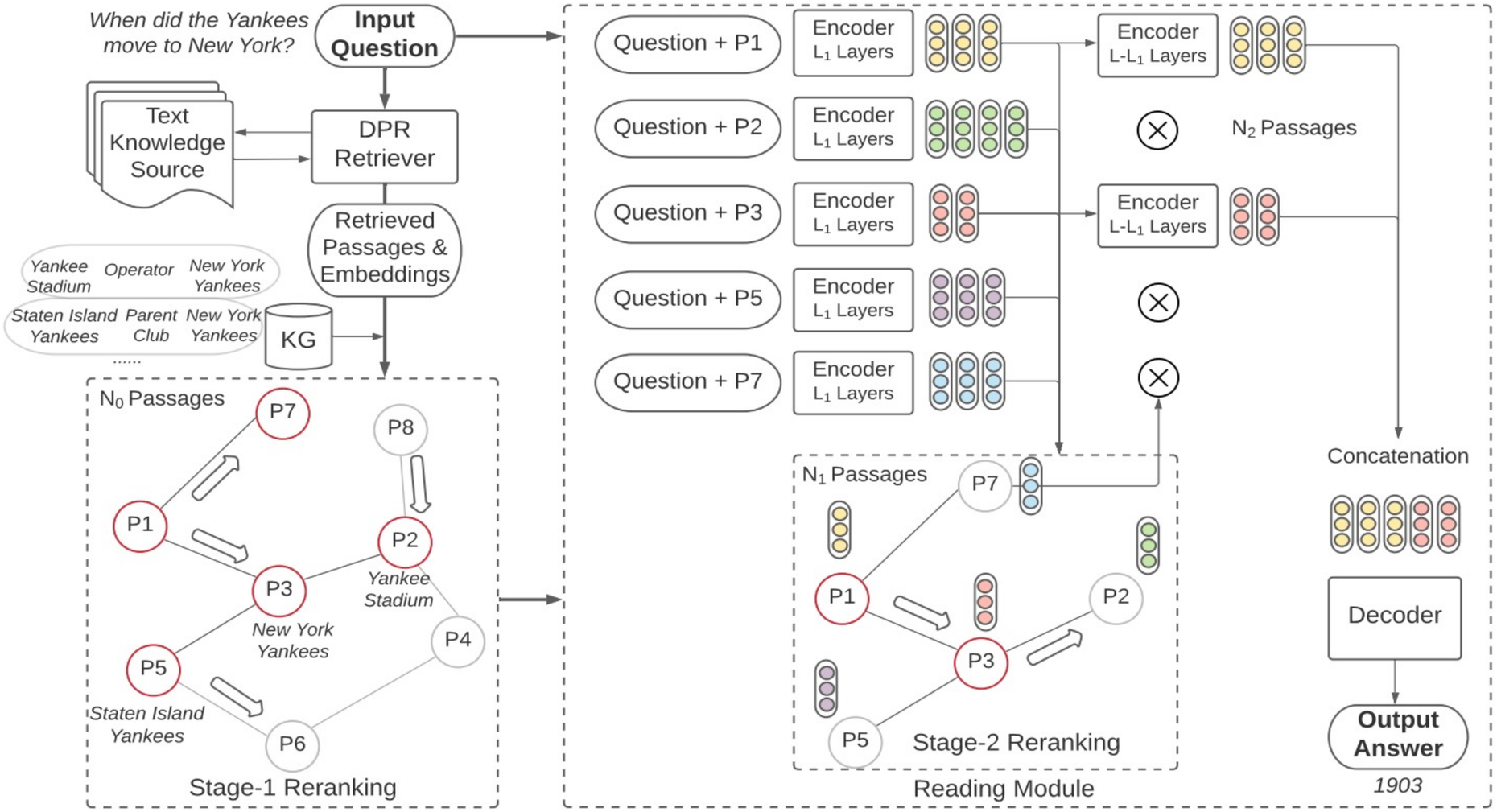
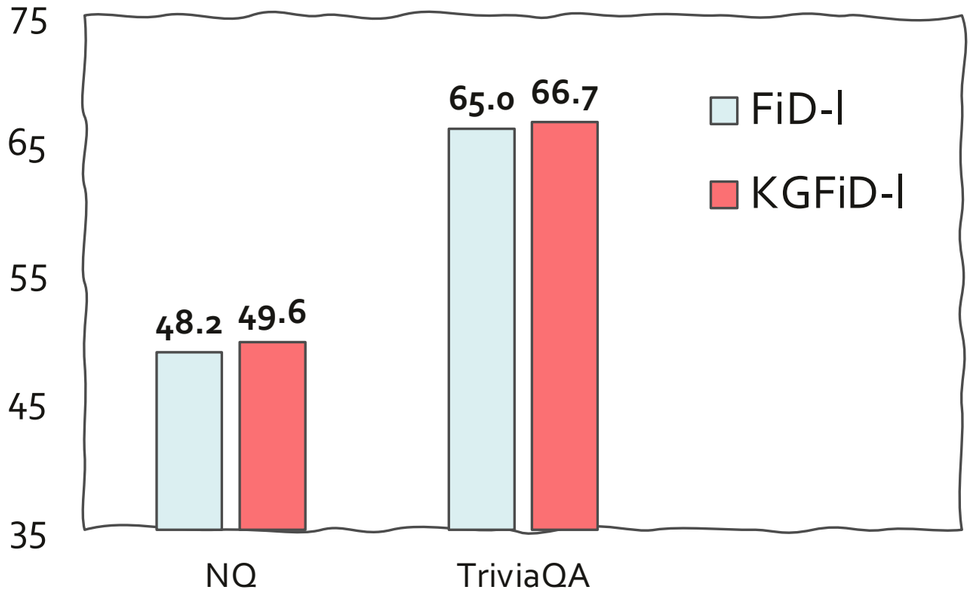


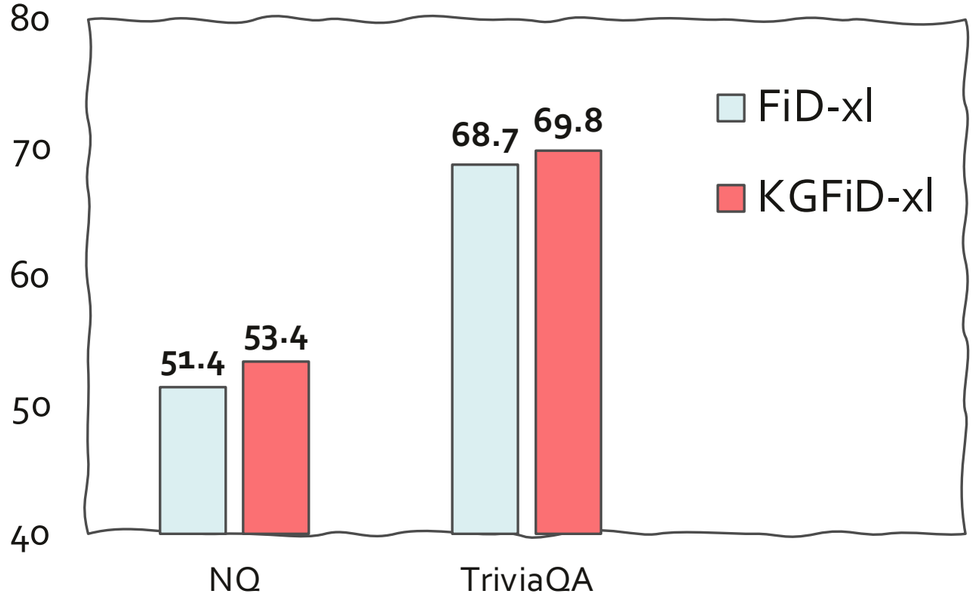
Figure: Overall Model Framework. Figure from KG-FiD paper figure 1.

[1] KG-FiD: Infusing Knowledge Graph in Fusion-in-Decoder for Open-Domain Question Answering. ACL 2022

KG-FiD: Downstream QA performance



KG-FiD vs. FiD (200M paras) +1.2%



KG-FiD vs. FiD (800M paras) +1.6%

Knowledge Graph Enhanced Passage Reader for Open-domain Question Answering

The 2022 Conference on Empirical Methods in Natural Language Processing

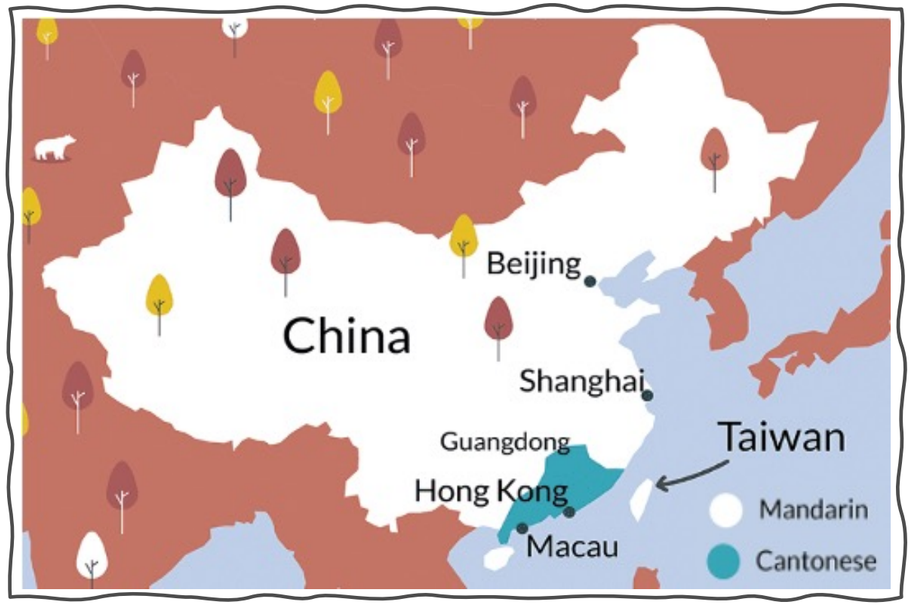
Wenhao Yu*, Mingxuan Ju* (equal contribution),
Tong Zhao, Chuxu Zhang, Yanfang Ye

Grape: KG for improving factuality

-- Why using KG? Retrieval-augmented LMs often produce answers that contradict the facts, due to inaccurate understanding of the retrieved factual evidence.

What is the primary language of China?

Wikipedia (Mandarin Chinese): Mandarin Chinese now dominates public life in mainland China. The only varieties of Chinese commonly taught in university courses are Mandarin and Cantonese.



SoTA output: Cantonese

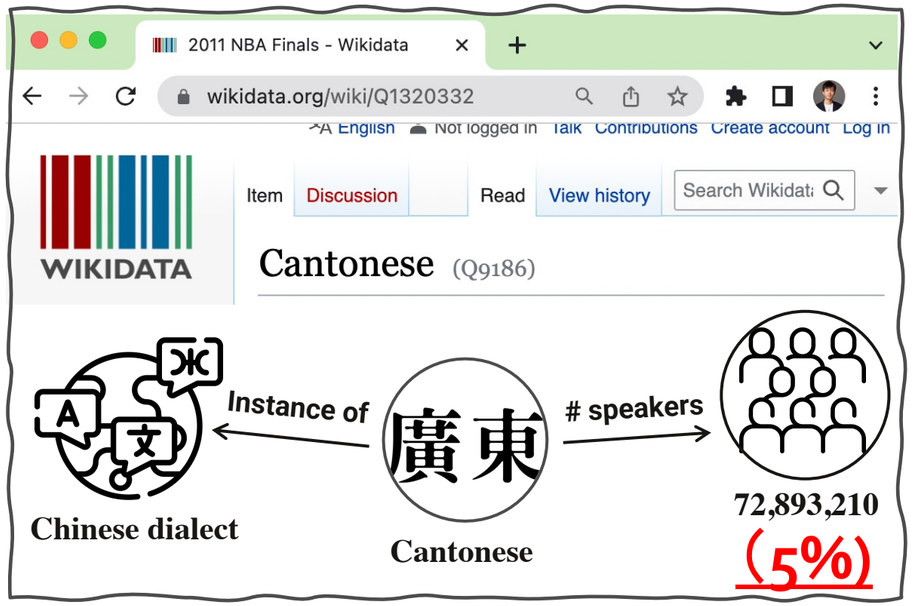
[1] Grape: Knowledge Graph Enhanced Passage Reader for Open-domain Question Answering. EMNLP 2022

Grape: KG for improving factuality

-- Why using KG? Retrieval-augmented LMs often produce answers that contradict the facts, due to inaccurate understanding of the retrieved factual evidence.

What is the primary language of China?

Wikipedia (Mandarin Chinese): Mandarin Chinese now dominates public life in mainland China. The only varieties of Chinese commonly taught in university courses are Mandarin and Cantonese.

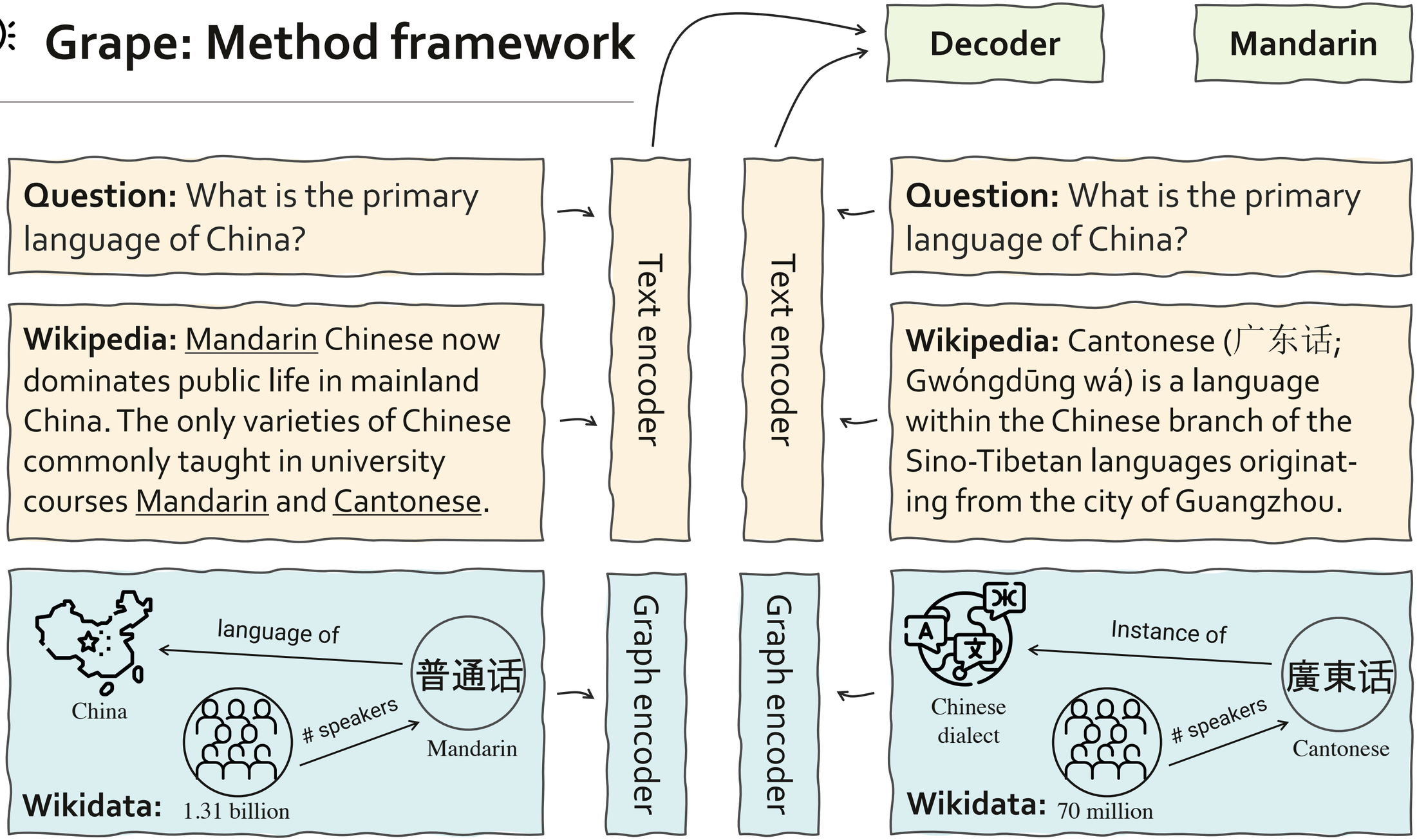


SoTA output: Cantonese ✗

Our output: Mandarin ✓

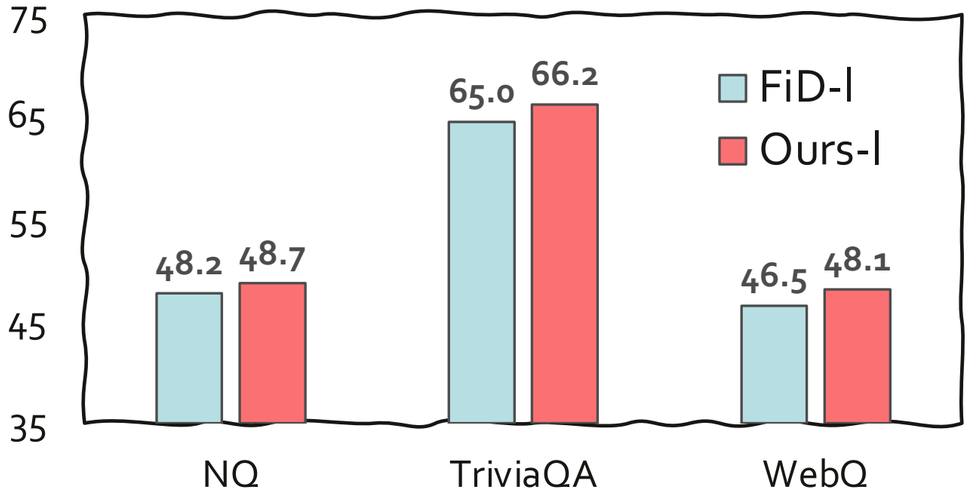
[1] Grape: Knowledge Graph Enhanced Passage Reader for Open-domain Question Answering. EMNLP 2022

Grape: Method framework

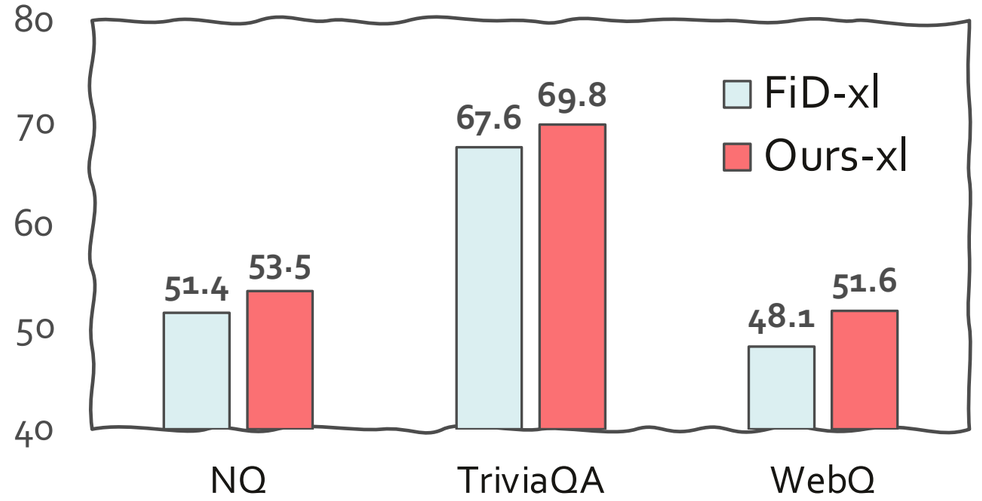


[1] Grape: Knowledge Graph Enhanced Passage Reader for Open-domain Question Answering. EMNLP 2022

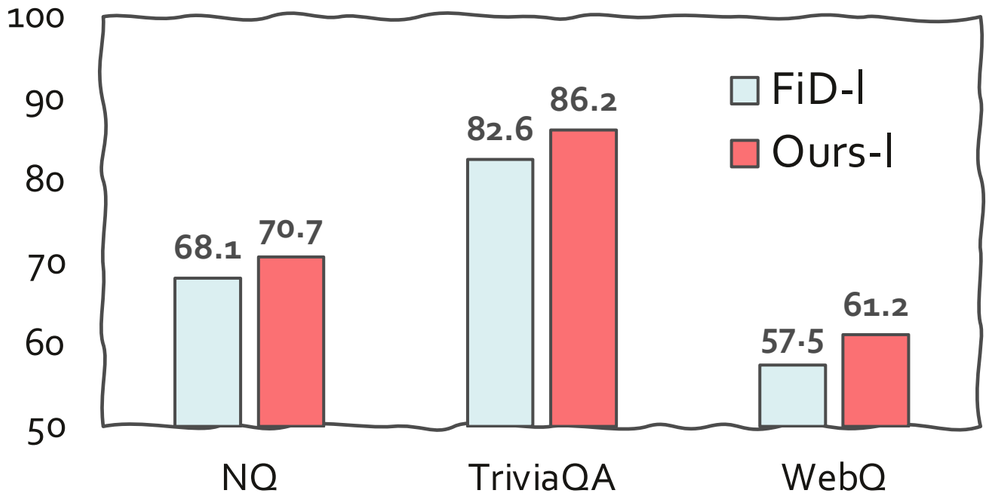
Grape: Downstream QA performance



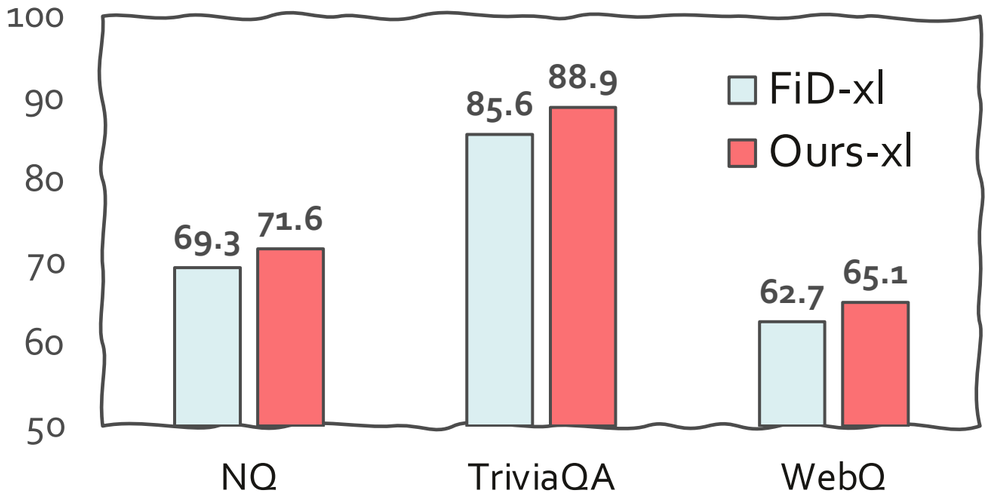
Grape vs. SoTA (200M paras) +1.1% full



Grape vs. SoTA (800M paras) +2.5% full



Grape vs. SoTA (200M) +3.3% 1-hop sub



Grape vs. SoTA (800M) +2.7% 1-hop sub

[1] Grape: Knowledge Graph Enhanced Passage Reader for Open-domain Question Answering. EMNLP 2022

Takeaway and future work

KG-FiD is the first work using the KG to improve passage retrieval via reranking.

GraPE is the first work using the KG (docs are fixed) to improve passage reader.

Dataset	FiD	KG-FiD	GraPE
NQ	51.4	53.4 (+2.0)	53.5 (+2.1)
TriviaQA	67.6	69.8 (+2.2)	69.8 (+2.2)

Can combine KG in retrieval + KG in reader leading to better performance?

Papers list

-- Retrieval augmented NLG model:

[1] Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. Neurips 2020

[2] Leveraging passage retrieval with generative models for open domain question answering. EACL 2021

-- Retrieval augmented NLG model + LLM (e.g., GPT-3):

[3] Generate rather than Retrieve: Large Language Models are Strong Context Generators. ICLR 2023

[4] REPLUG: Retrieval-Augmented Black-Box Language Models. ArXiv 2023

-- Knowledge graph augmented NLG model:

[5] KG-FiD: Infusing Knowledge Graph in Fusion-in-Decoder for Open-Domain Question Answering. ACL 2022

[6] Grape: Knowledge Graph Enhanced Passage Reader for Open-domain Question Answering. EMNLP 2022

- -Memory augmented NLG model:

[7] A Unified Encoder-Decoder Framework with Entity Memory. EMNLP 2022

-- Knowledge augmented NLG model applications:

[8] Diversifying Content Generation with Mixture of Knowledge Graph Experts. ACL 2022

[9] Retrieval Augmentation for Commonsense Reasoning: A Unified Approach. EMNLP 2022

[10] Retrieval-Augmented Multimodal Language Modeling. ArXiv 2022

A Unified Encoder-Decoder Framework with Entity Memory

The 2022 Conference on Empirical Methods in Natural Language Processing

Zhihan Zhang, Wenhao Yu, Chenguang Zhu, Meng Jiang

EDMem: Entity memory as knowledge

-- Baseline 1: close-book

Q: At the equator, what is the speed of the ground as a result of the Earth's rotation, in miles per hour?

A: 8,000

Lack of entity knowledge!

-- Baseline 2: open-book

Q: At the equator, what is the speed of the ground as a result of the Earth's rotation, in miles per hour?

Wikipedia: At the equator, the circum-ference of the Earth is 40,070 kilometers, and the day is 24 hours long so the speed is 1670 kilometers/hour (**1037 miles/hr**)

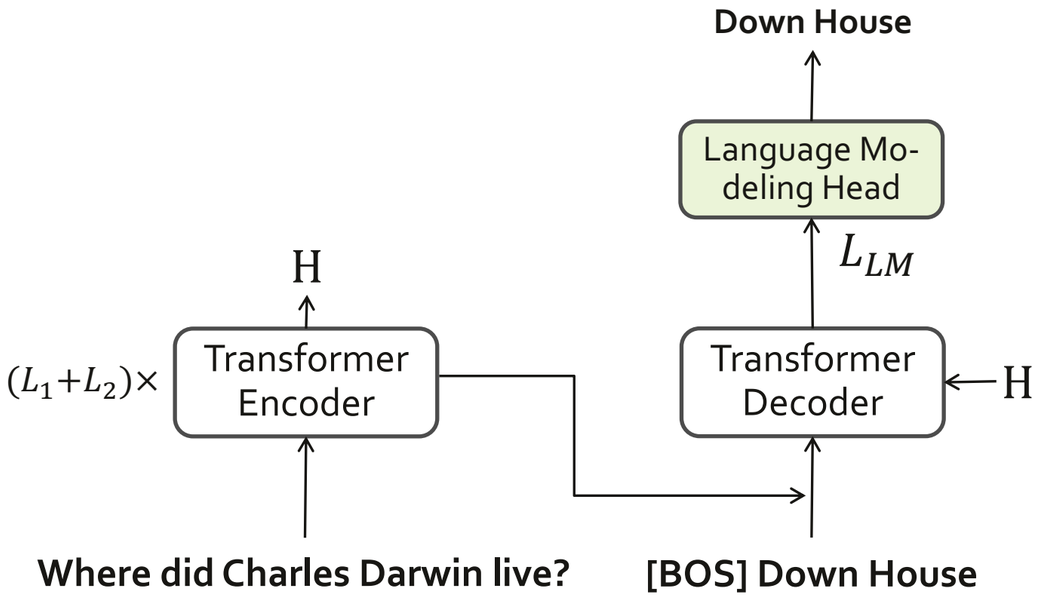
A: 1037 mph

Lack efficiency! (retrieve + read)

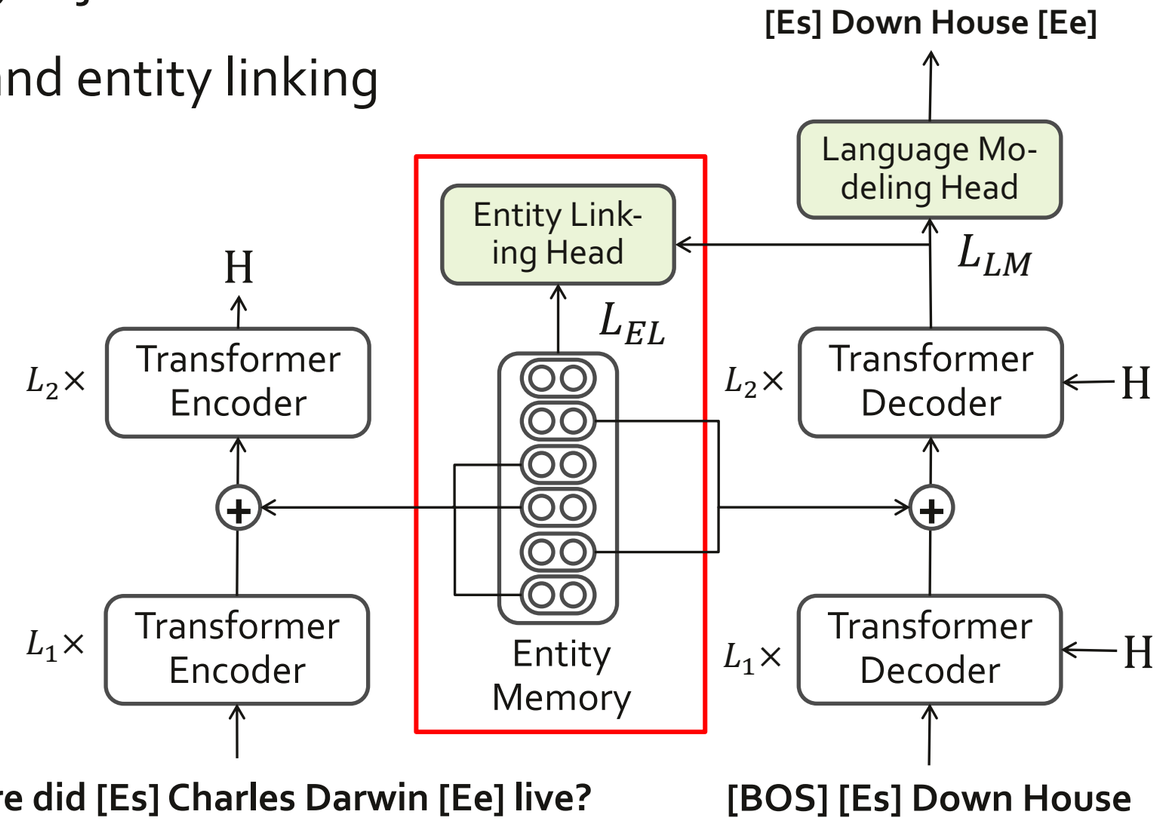
[1] A Unified Encoder-Decoder Framework with Entity Memory. EMNLP 2022

EDMem: Proposed Method -- Training

- EDMem is based on an encoder-decoder Transformer framework
- Pre-trained on Wikipedia passages. Pre-training objective includes random token masking, salient span masking, and entity linking



(a) Traditional encoder-decoder framework



(b) Encoder-decoder with entity memory

[1] A Unified Encoder-Decoder Framework with Entity Memory. EMNLP 2022

EDMem: Proposed Method -- Inference

(1) Free-form generation

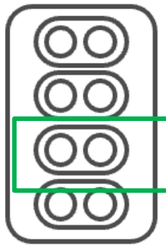
Where is Niagara?

Niagara is in Niagara County, Pennsylvania ✗, United States ⚠ Invalid entity name!

Entity-Aware Decoding

Constrained generation on selected entities from the memory

Niagara is in Niagara County, New York ✓, United States



(2) Static entity linking:

Predict top-k entities in the 1st run, then use these entities as constraints in the 2nd run

(3) Dynamic entity linking:

Predict top-k entities on-the-fly for constrained decoding in the same run

[1] A Unified Encoder-Decoder Framework with Entity Memory. EMNLP 2022

EDMem: Proposed Method -- Inference

Method	Open-domain QA			Language Generation			Inference time on TriviaQA
	TriviaQA	NQ	WebQ	MSMARCO	CREAK	ELI5	
Close-book Methods							
BART-large	25.02	24.82	29.33	53.26	30.34	22.99	17 s
EDMem (1) Free form	42.24	29.14	36.47	**54.45**	**31.78**	**23.24**	28 s
EDMem (2) Static	**46.19**	**30.19**	**41.44**	-	-	-	59 s
EDMem (3) Dynamic	43.82	27.70	39.52	52.84	30.68	23.97	48 s
Open-book Methods							
FiD (SoTA)	67.60	51.40	47.64	57.22	33.54	22.12	41 min

EDMem achieved much better accuracy over BART, better efficiency over open-book methods

[1] A Unified Encoder-Decoder Framework with Entity Memory. EMNLP 2022

Papers list

-- Retrieval augmented NLG model:

[1] Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. Neurips 2020

[2] Leveraging passage retrieval with generative models for open domain question answering. EACL 2021

-- Retrieval augmented NLG model + LLM (e.g., GPT-3):

[3] Generate rather than Retrieve: Large Language Models are Strong Context Generators. ICLR 2023

[4] REPLUG: Retrieval-Augmented Black-Box Language Models. ArXiv 2023

-- Knowledge graph augmented NLG model:

[5] KG-FiD: Infusing Knowledge Graph in Fusion-in-Decoder for Open-Domain Question Answering. ACL 2022

[6] Grape: Knowledge Graph Enhanced Passage Reader for Open-domain Question Answering. EMNLP 2022

- -Memory augmented NLG model:

[7] A Unified Encoder-Decoder Framework with Entity Memory. EMNLP 2022

-- Knowledge augmented NLG model applications:

[8] Diversifying Content Generation with Mixture of Knowledge Graph Experts. ACL 2022

[9] Retrieval Augmentation for Commonsense Reasoning: A Unified Approach. EMNLP 2022

[10] Retrieval-Augmented Multimodal Language Modeling. ArXiv 2022

Diversifying Content Generation for Commonsense Reasoning with Mixture of Knowledge Graph Experts

The 2022 Annual Meeting of the Association for Computational Linguistics

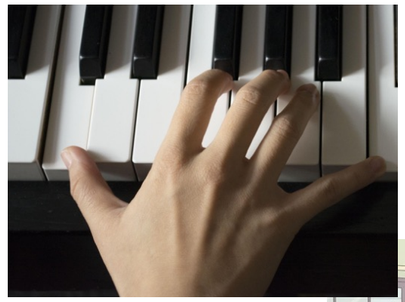
Wenhao Yu, Chenguang Zhu, Lianhui Qin, Zhihan Zhang,
Tong Zhao, Meng Jiang

MoKGE: Knowledge enhanced diverse generation

-- Example of diverse generation in ComVE dataset

- **Input:** Piano is a kind of sport.
  (a counterfactual statement)

- **Output 1:** You can produce music when pressing keys on the piano, so it is an instrument. **(usage)**



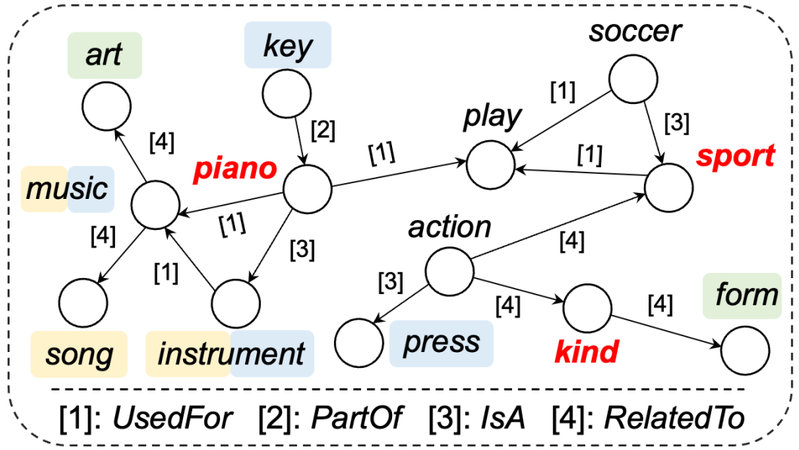
- **Output 2:** Piano is a musical instrument used in songs to produce different musical tones. **(effect)**

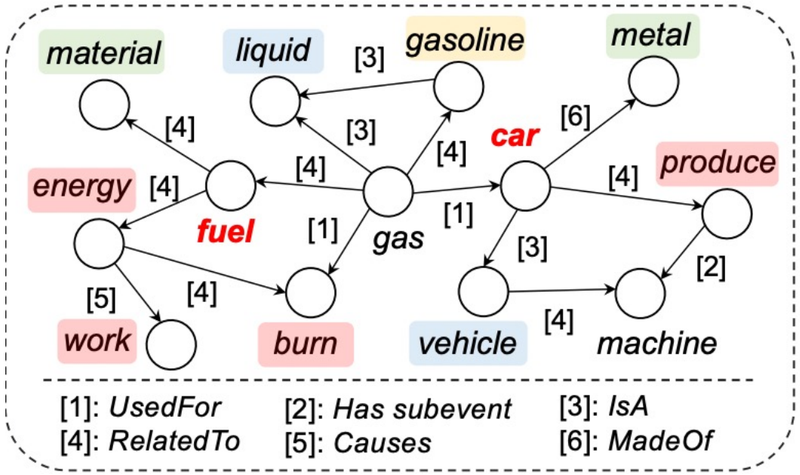
- **Output 3:** Piano is a kind of art form. **(taxonomy)**

Many plausible reasons from different perspectives!

[1] Diversifying Content Generation for Commonsense Reasoning with Mixture of Knowledge Graph Experts. ACL 2022

MoKGE: Knowledge expert

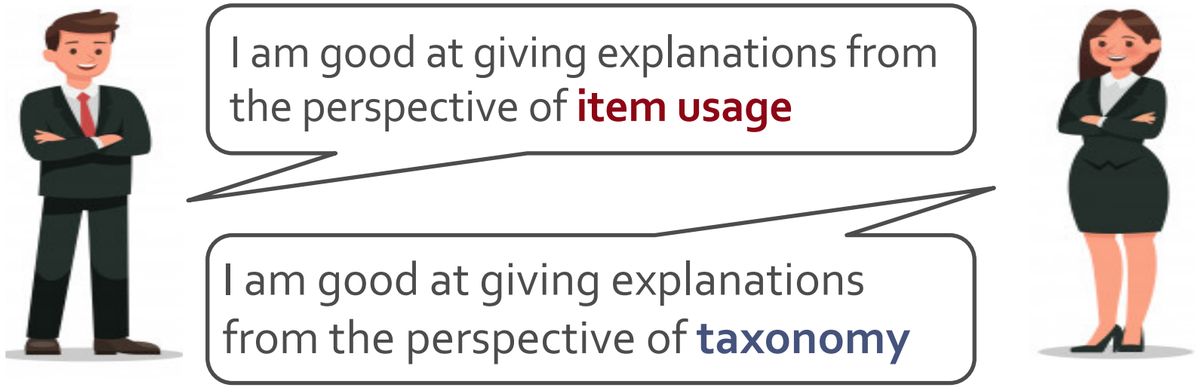




-- Input: **Piano** is a kind of **sport** .

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** .



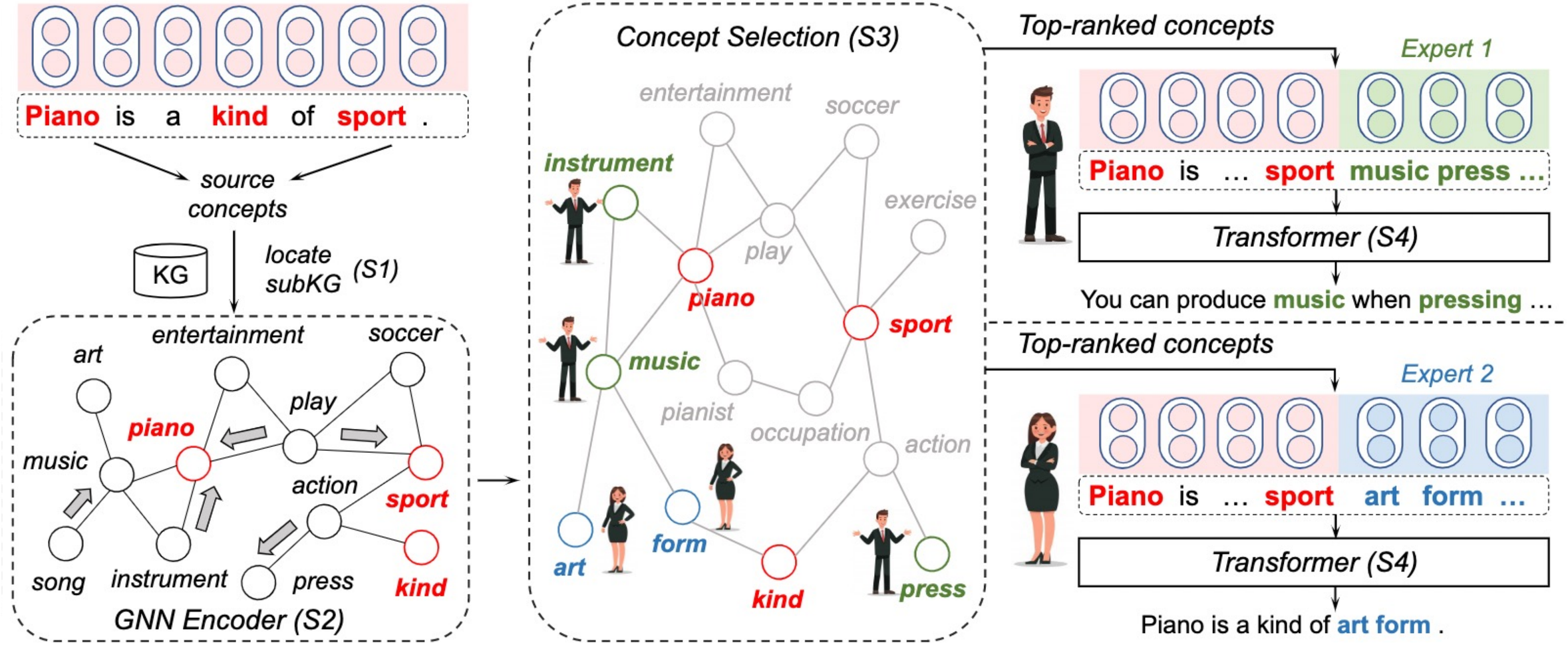
-- Input: **Cars** are made of **fuel**.

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

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

[1] Diversifying Content Generation for Commonsense Reasoning with Mixture of Knowledge Graph Experts. ACL 2022

MoKGE: Model framework



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

[1] Diversifying Content Generation for Commonsense Reasoning with Mixture of Knowledge Graph Experts. ACL 2022

MoKGE: Downstream performance

-- Automatic evaluation:

ComVE

Method	Diversity Metric		Quality Metric
	Self-BLUE-4 (↓)	Distinct-2 (↑)	BLUE-4 (↑)
SoTA -- (MoE)	33.42	46.93	18.91
MoKGE (ours)	**30.93**	**48.44**	**19.13**

α-NLG

Method	Diversity Metric		Quality Metric
	Self-BLUE-4 (↓)	Distinct-2 (↑)	BLUE-4 (↑)
SoTA -- (MoE)	24.04	36.22	14.31
MoKGE (ours)	**22.43**	**38.15**	13.74

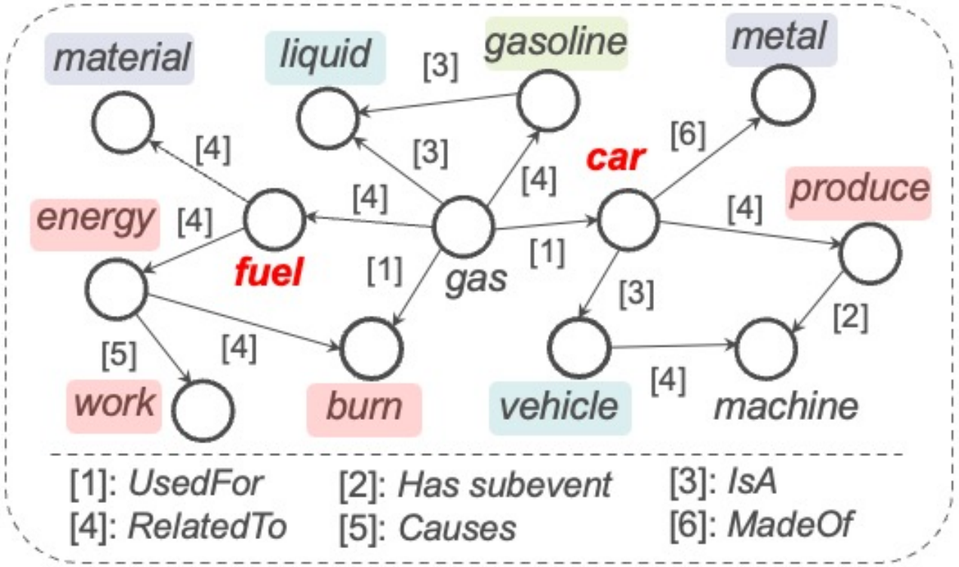
MoKGE (ours) performs better than MoE (SoTA) on diversity, and on par on quality.

[1] Diversifying Content Generation for Commonsense Reasoning with Mixture of Knowledge Graph Experts. ACL 2022

MoKGE: Downstream performance

-- Case Study:

Input: Cars are made of fuel
Output:
(1) Cars are not made of fuel.
(2) Cars burn fuel to produce energy and work.
(3) Fuel is a liquid which cannot make cars.



MoE:
(1) Cars are made of rubber. Fuel is not used to make cars.
(2) Cars are made of aluminum, which is not fuel.
(3) Cars are powered by electric motors and not by fuel.

MoKGE:
(1) Fuel is not a vehicle material.
(2) Fuel is not used to make cars. They use gasoline.
(3) Cars are not made of fuel. They are made of metal.

[1] Diversifying Content Generation for Commonsense Reasoning with Mixture of Knowledge Graph Experts. ACL 2022

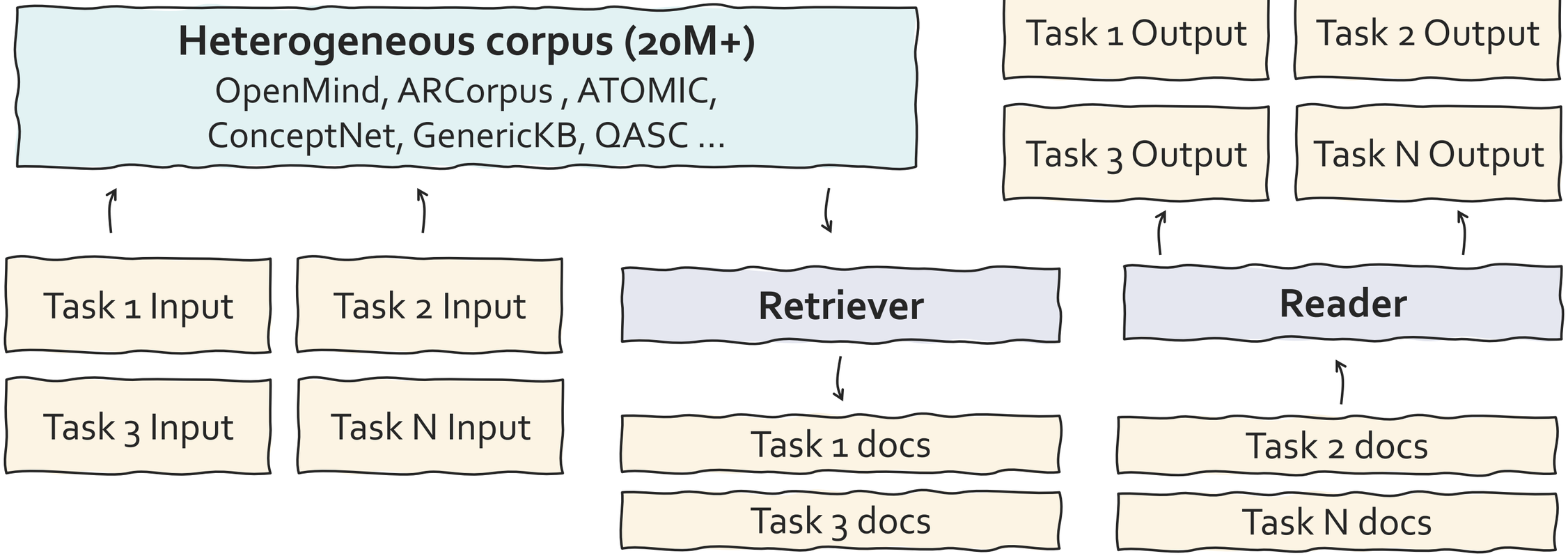
Retrieval Augmentation for Commonsense Reasoning: A Unified Approach

The 2022 Conference on Empirical Methods in Natural Language Processing

Wenhao Yu, Chenguang Zhu, Zhihan Zhang, Shuohang Wang,
Zhuosheng Zhang, Yuwei Fang, Meng Jiang

RACo: What is a unified approach?

-- The goal of a unified approach in our paper is to use only **one corpus** and **one retriever** to solve various kind of commonsense reasoning tasks.



[1] Retrieval Augmentation for Commonsense Reasoning: A Unified Approach. EMNLP 2022

RACo: Why unified corpus?

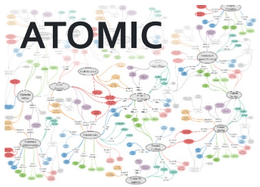
-- **World knowledge is heterogeneous!** Knowledge cannot be contained in a single (unstructured) document corpus or (structured) knowledge graph.

(Binary QA): Harry Potter can teach classes on how to fly on a broomstick.

Harry Potter is a wizard … Harry plays Quidditch while riding on a broomstick.

OPEN MIND

Harry Potter is good at riding broomstick.



Someone who's good at something can teach it.

[1] Retrieval Augmentation for Commonsense Reasoning: A Unified Approach. EMNLP 2022

RACo: Why unified retriever?

-- Training retrievers on each dataset is expensive, and easily overfits! And ground truth (Q, D) pair is not always available when training a retriever.

Query embedding

Document embedding

CSQA: Only after learning you can gain knowledge. (Yes or No)

Baseline[1]

Human: Learning causes you to gain knowledge

ComVE: The sun made my t-shirt wet. (Explanation generation)

Ours (1)

Generative model: Knowledge is very import-ant, and you should learn as much as you can.

Ours (2)

Golden label: Knowledge is very import-ant, and you should learn as much as you can.

Multi-tasks, multi-datasets retriever training!

[1] Retrieval Augmentation for Commonsense Reasoning: A Unified Approach. EMNLP 2022

RACo: Proposed corpus and method

(1) We collected and publicized a collection of over 20 million documents from 19 knowledge sources for commonsense knowledge retrieval.

(2) We proposed novel strategies (multi-task, multi-datasets) for training a unified strong commonsense knowledge retriever.

(3) We achieved new SoTA on CommonGen and CREAK leaderboards.

CommonGen Leaderboard (V1.1)

Rank	Model	BLEU-4	CIDEr	SPICE
	Upper Bound	46.49	37.64	52.43
1 May 15, 2022	**RACo** *Microsoft Cognitive Services Research Group* Email Pending	43.12	19.144	34.028

Book: In-Domain

	ystem	Dev Set	Test Set	Contrastive Set
	uman (ensembled)	99.0	-	99.0
	uman (averaged)	96.3	-	92.2
1	RACo-Large	88.2	88.6	74.4
2	T5-3B	85.6	85.1	70.0

- CommonGen: https://inklab.usc.edu/CommonGen/leaderboard.html
- CREAK: https://www.cs.utexas.edu/~yasumasa/creak/leaderboard.html

[1] Retrieval Augmentation for Commonsense Reasoning: A Unified Approach. EMNLP 2022

RACo: Downstream task performance

Multi-choice Commonsense Question Answering

(Accuracy)	CSQA1.0	OBQA
RACo	**75.76**	**71.25**
GreaseLM	74.20	66.90
QA-GNN	73.40	67.80
UNICORN	71.60	70.02

Commonsense Fact Verification

(Accuracy)	CSQA2.0	CREAK
RACo	**61.75**	**84.17**
UNICORN	54.90	79.51
GreaseLM	-	77.51
T5-Large	54.60	77.32

Constrained Commonsense Generation

CommonGen	(BL-4)	(SPICE)
RACo	**42.76**	**33.89**
KFCNet	41.97	33.11
UNICORN	39.86	33.20
KG-BART	30.90	32.70
CALM	29.50	30.20

Counterfactual Explanation Generation

ComVE	(BL-4)	(SPICE)
RACo	**25.30**	**36.37**
UNICORN	24.46	35.79
CALM	23.50	35.23
MoKGE	22.87	34.88
T5-Large	22.77	34.62

[1] Retrieval Augmentation for Commonsense Reasoning: A Unified Approach. EMNLP 2022

Retrieval-Augmented Multimodal Language Modeling

Michihiro Yasunaga, Armen Aghajanyan, Weijia Shi, Rich James, Jure Leskovec, Percy Liang, Mike Lewis, Luke Zettlemoyer, Wen-tau Yih

Retrieval-Augmented Multimodal Language Modeling

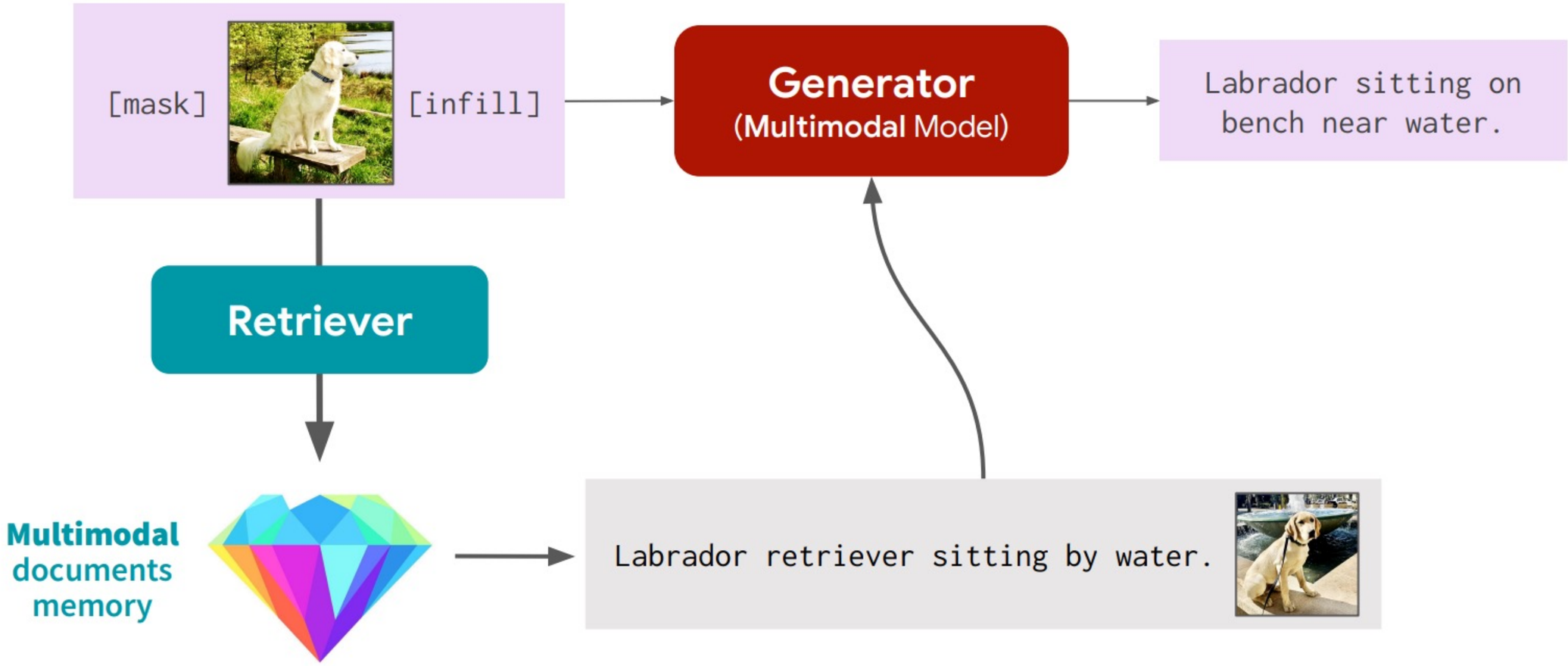
Text to image

Image to text

The Dragon and Tiger Pagodas next to fireworks.

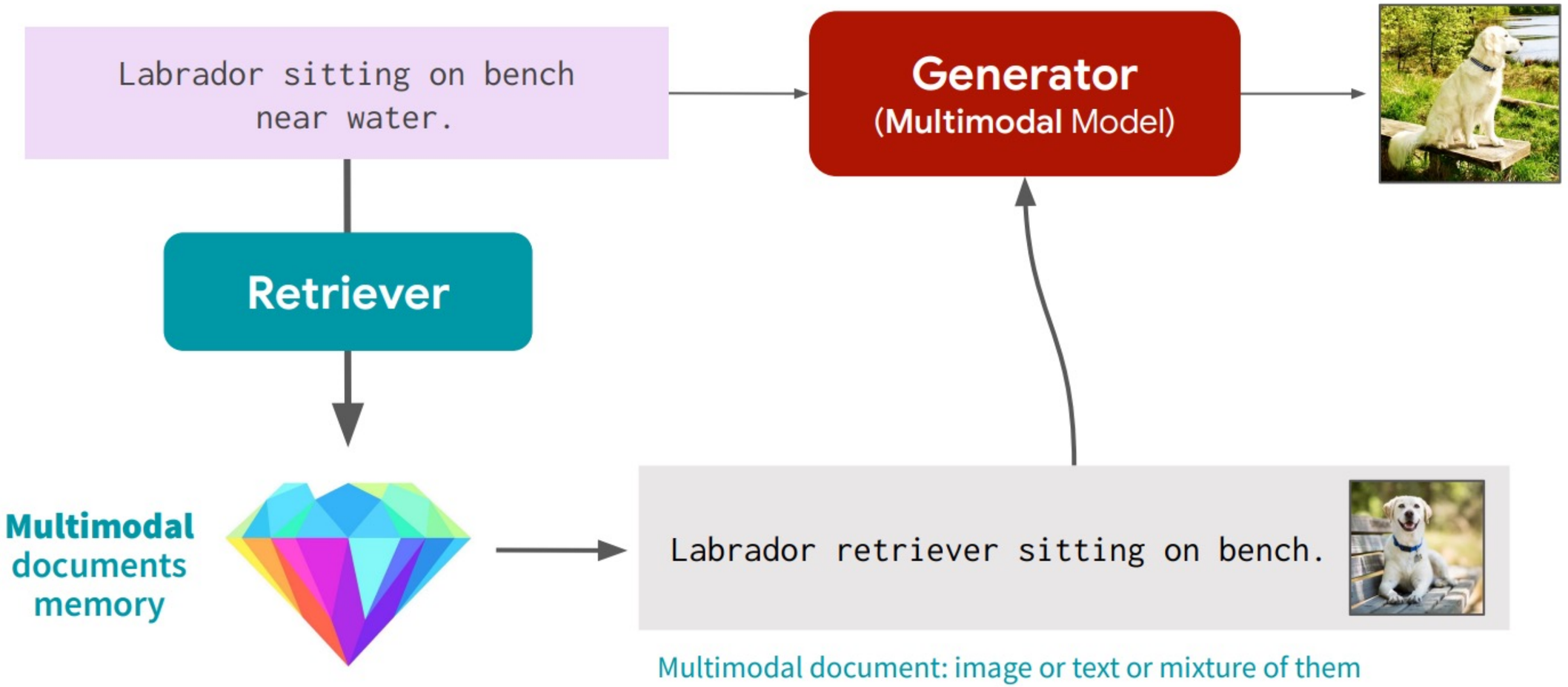
[1] Retrieval-Augmented Multimodal Language Modeling. ArXiv 2022

RA-CM3: Image-to-text



[1] Retrieval-Augmented Multimodal Language Modeling. ArXiv 2022

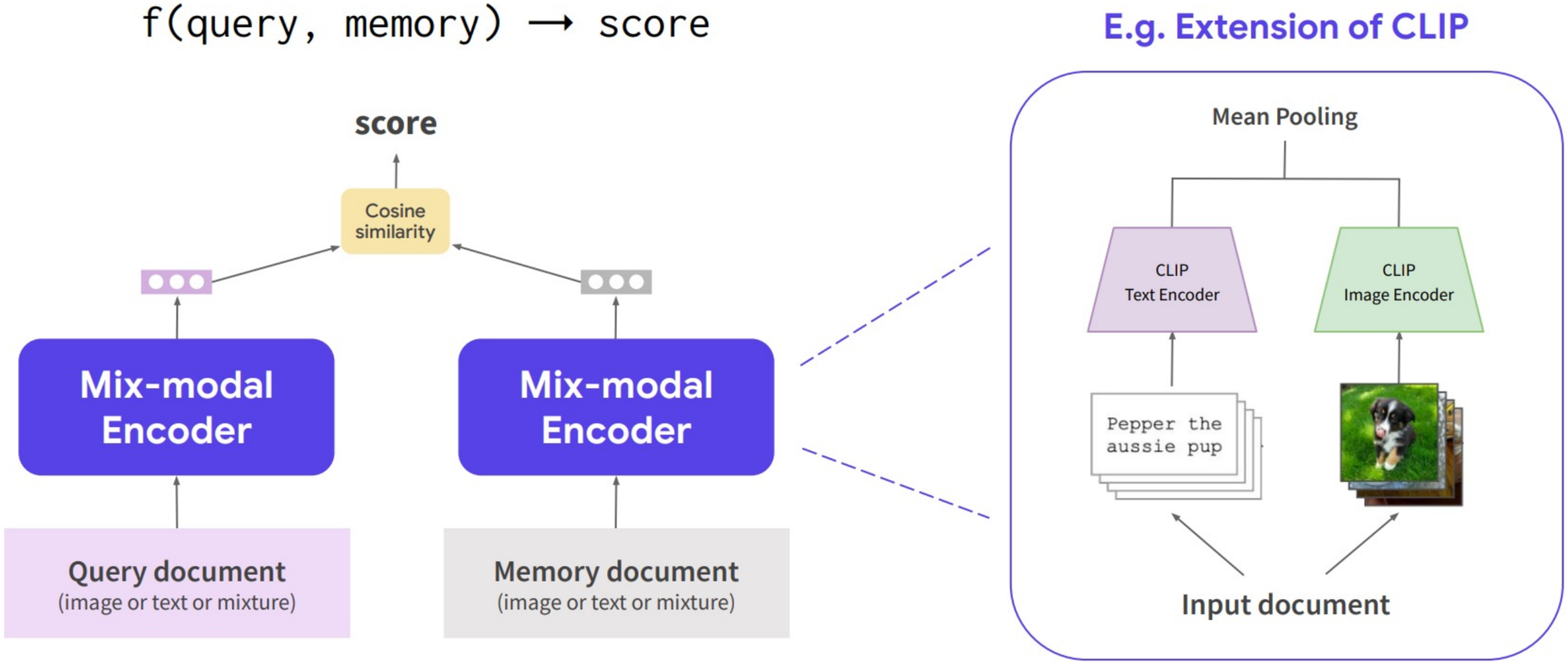
RA-CM3: Text-to-Image



[1] Retrieval-Augmented Multimodal Language Modeling. ArXiv 2022

RA-CM3: Multi-modal retriever

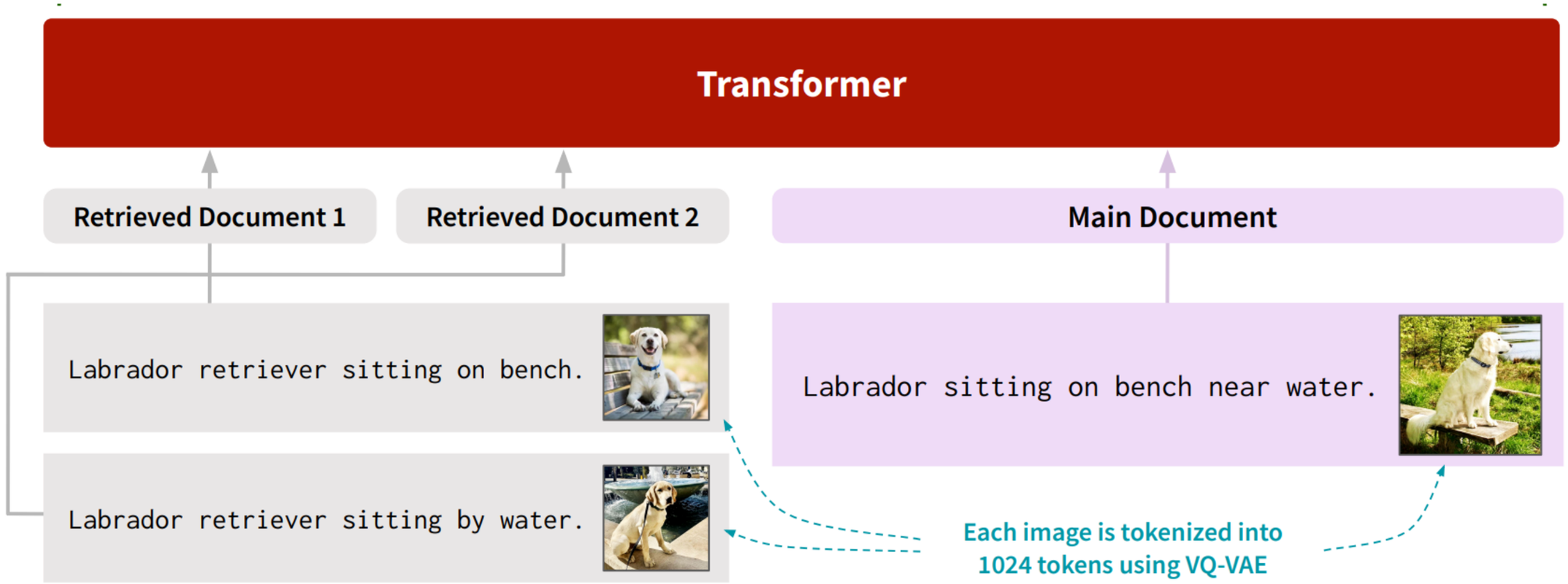
-- The multimodal retriever is a dense retriever with a mixed-modal encoder that can encode mixture of text and images



[1] Retrieval-Augmented Multimodal Language Modeling. ArXiv 2022

RA-CM3: Multi-modal generator

-- The generator uses the CM3 Transformer architecture, and we prepend the retrieved documents to the main input document that we feed into the model.



[1] Retrieval-Augmented Multimodal Language Modeling. ArXiv 2022

RA-CM3: Experiments

RA-CM3 In-context

Armenian church

RA-CM3 outputs

Baseline outputs

(Vanilla CM3)

(Stable Diffusion)

An **Armenian church** during a sunny day.

Mount Rainier

People standing in front of the **Mount Rainier**.

[1] Retrieval-Augmented Multimodal Language Modeling. ArXiv 2022

RA-CM3: Experiments

RA-CM3 In-context

Mount Rushmore

Japanese cherry

RA-CM3 outputs

Baseline outputs

(Vanilla CM3)

(Stable Diffusion)

The **Mount Rushmore** with **Japanese cherry** trees in the front.

Oriental Pearl tower

The Oriental Pearl tower in oil painting.

[1] Retrieval-Augmented Multimodal Language Modeling. ArXiv 2022

Wrap-up

Three knowledge-augmented NLG directions

- Retrieval -> Structured (KG), unstructured (text retrieval) and Heterogeneous knowledge
- Memory network -> more accurate than close-book and more efficiency than open-book
- IR + LLM -> LLM generate + LLM read (GenRead) v.s. IR retrieve + LLM to read (RePLUG)
- Applciations: Diverse generation, Commonsense reasoning, Multi-modal learning

Survey paper:

- Wenhao Yu et al. *A survey of knowledge-enhanced text generation.* Accepted to ACM Computing Surveys, 2022.
- GitHub: https://github.com/wyu97/KENLG-Reading

Find us, when you go to the next conference!

Survey paper:

- Wenhao Yu et al. *A survey of knowledge-enhanced text generation.* Accepted to ACM Computing Surveys, 2022.
- GitHub: https://github.com/wyu97/KENLG-Reading

Survey paper:

- Wenhao Yu et al. *A survey of knowledge-enhanced text generation.* Accepted to ACM Computing Surveys, 2022.
- GitHub: https://github.com/wyu97/KENLG-Reading