Natural Language Processing

Berkeley
N L P

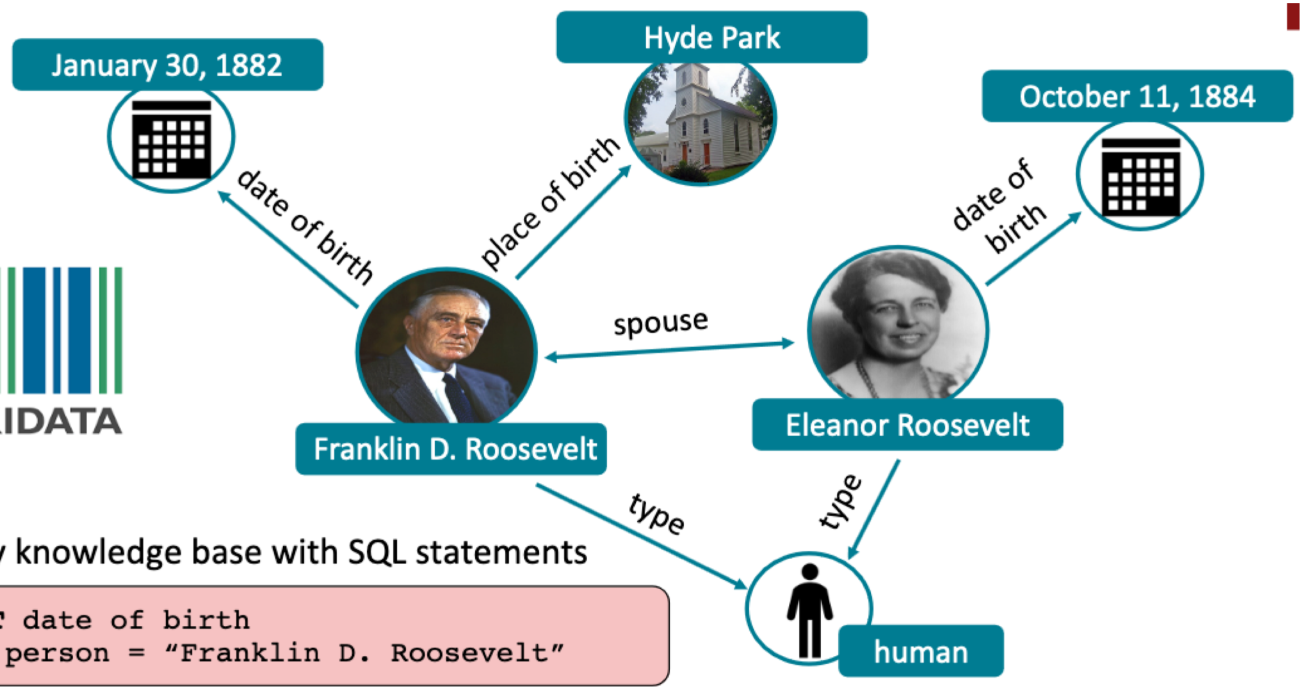
Retrieval / Knowledge-Intensive NLP

Kevin Lin – UC Berkeley

March 22, 2023

Retrieval / Knowledge-Intensive NLP

Structured Knowledge

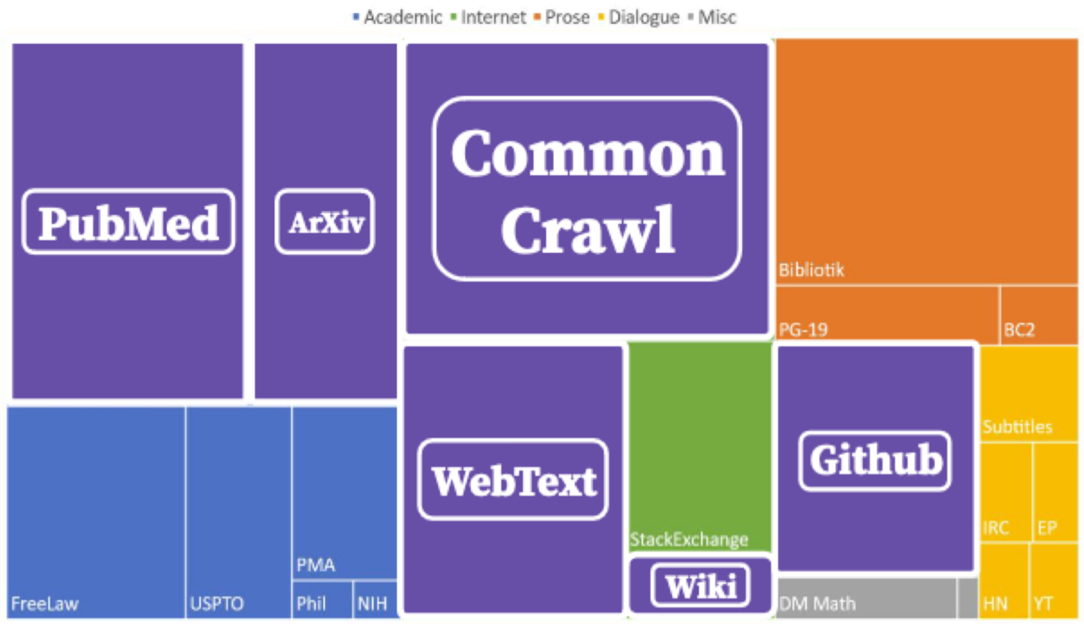


- Query knowledge base with SQL statements

```
SELECT date of birth
WHERE person = "Franklin D. Roosevelt"
```

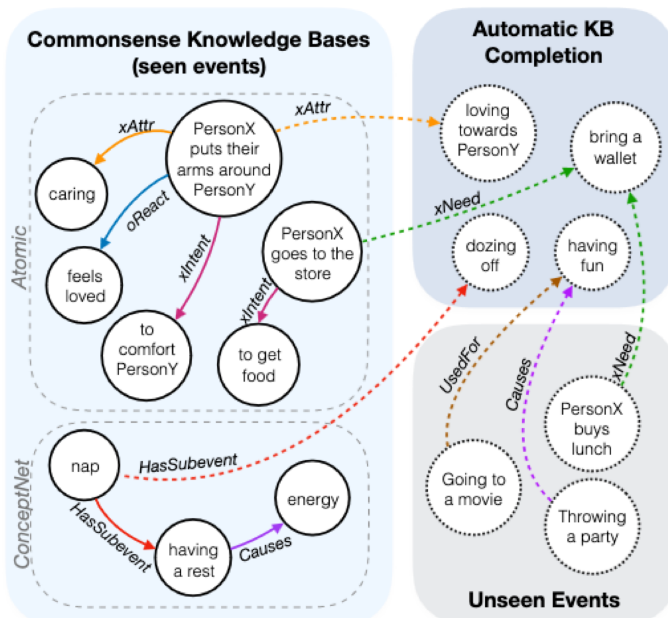
Unstructured Knowledge

- Large corpuses for LLMs contain lots of information / data
- The Pile (Gao et al., 2020)
  - 800GB

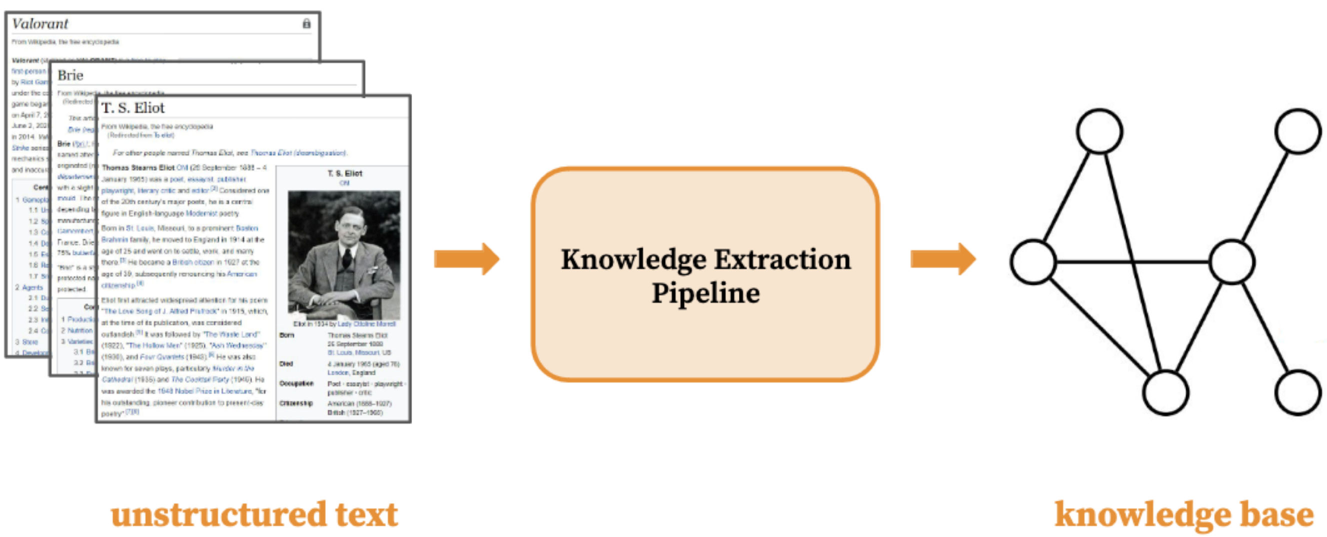


Semi-Structured Knowledge

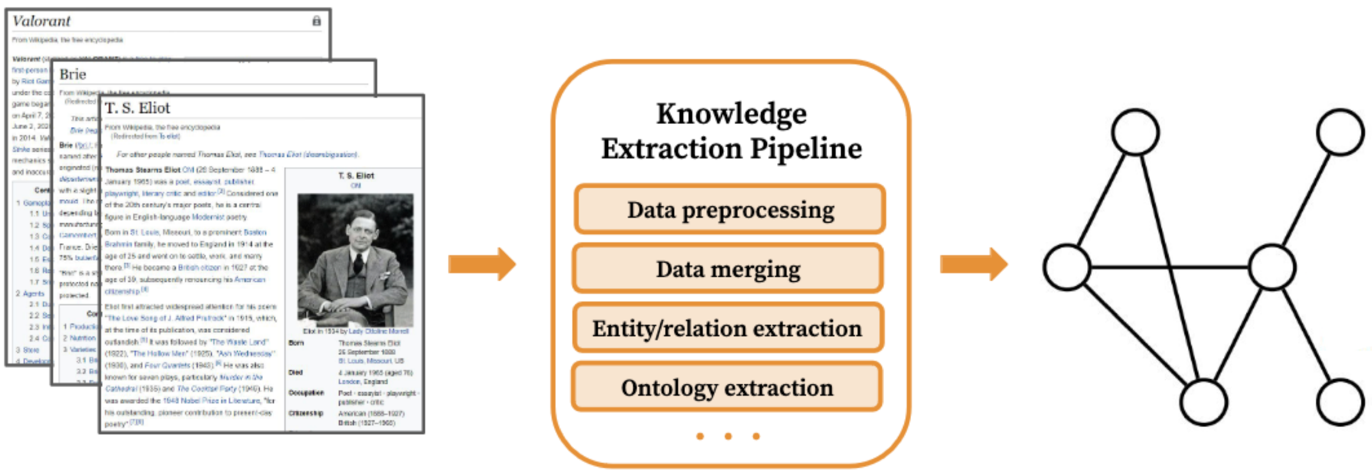
- Loosely structured KB with open-text
  - Eg. Common sense KBs



Structured Knowledge

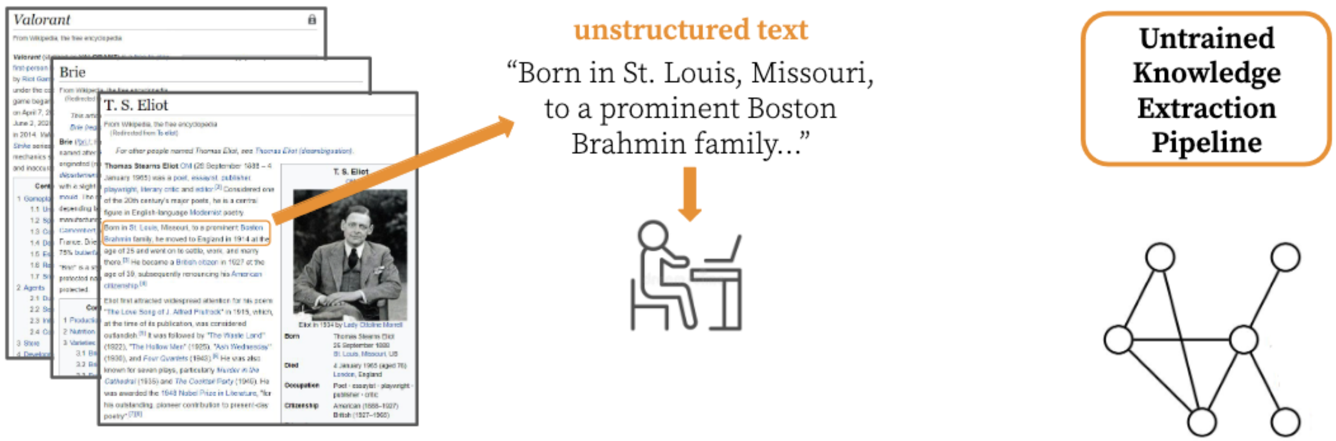


Structured Knowledge



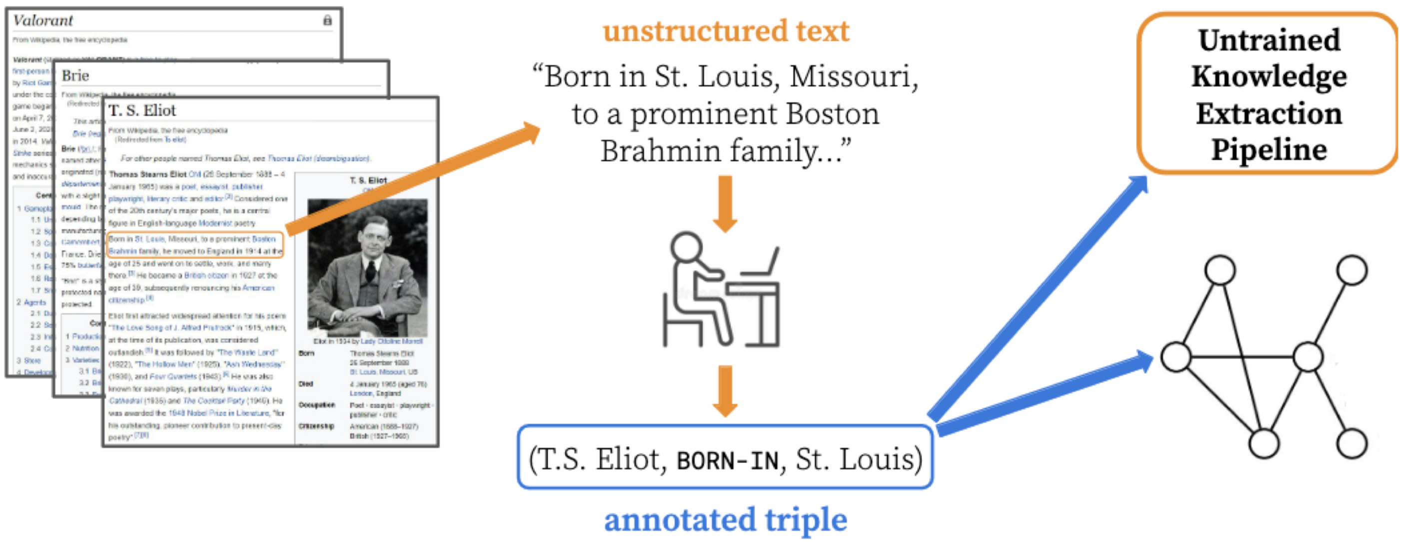
Populating the knowledge base often involves **complicated, multi-step NLP pipelines**

Structured Knowledge



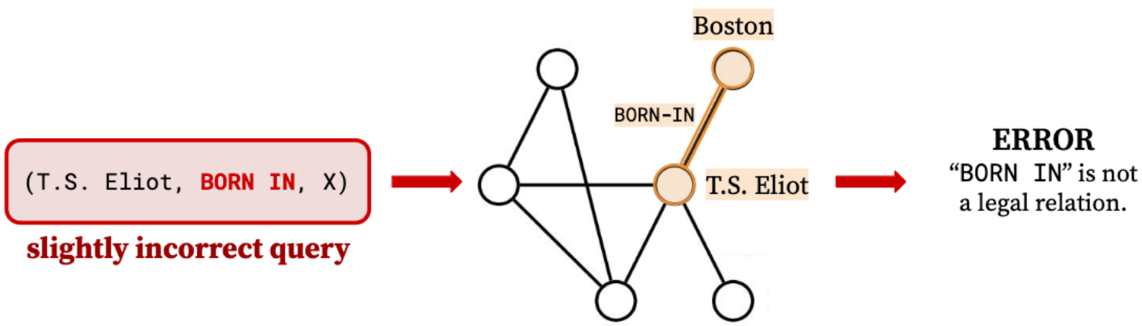
Requires **supervised data** to train the pipeline and/or fill the knowledge base

Knowledge Base Downsides



Requires **supervised data** to train the pipeline and/or fill the knowledge base

Structured Knowledge



Reliant on **fixed schemas** to store or query data

What do the LLMs “know?

- iPod Touch is produced by ___.
- London Jazz Festival is located in ____.
- Dani Alves plays with _____.
- Carl III used to communicate in ______.
- Ravens can _____.

What do the LLMs “know?

- iPod Touch is produced by Apple.
- London Jazz Festival is located in London.
- Dani Alves plays with Santos.
- Carl III used to communicate in German.
- Ravens can fly.

(Petroni et al., 2019)

What do the LLMs “know?

- Lots of knowledge from language modeling
- Issues:
  - Coverage: was the fact in the training set?
  - Frequency of facts: has the LM ”seen” it enough times?
  - Prompt sensitivity: if we reword it, will the answer change?

Language Models as Knowledge Bases?

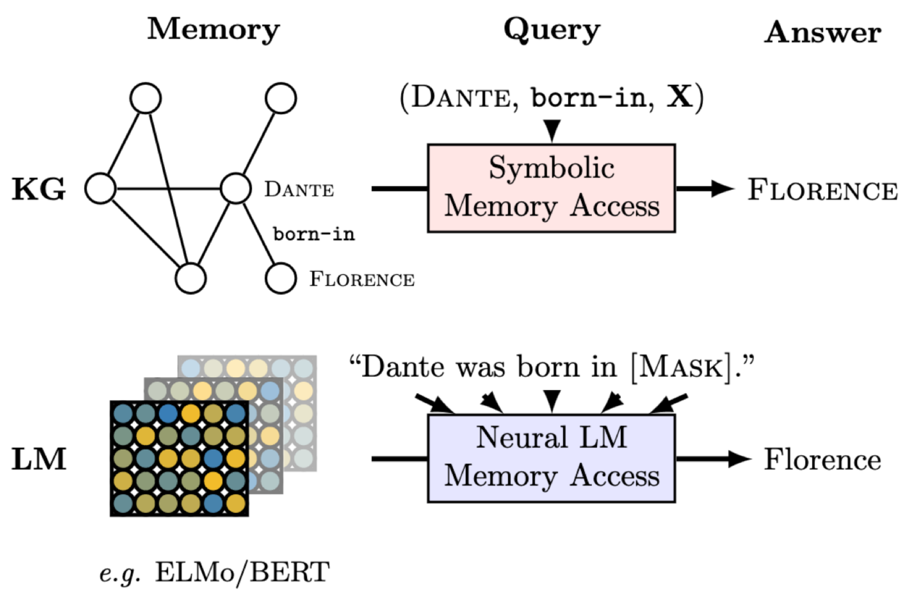
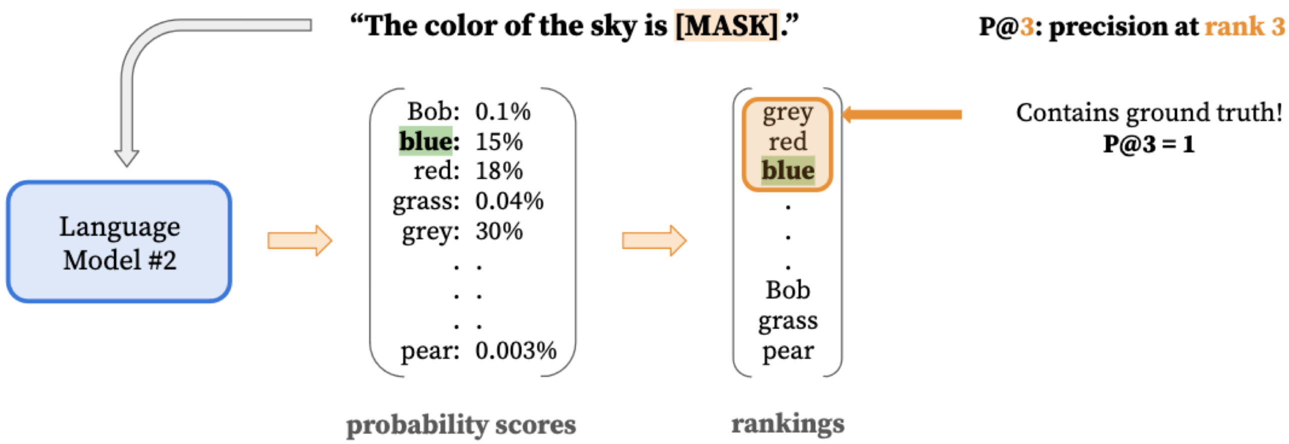


Figure 1: Querying knowledge bases (KB) and language models (LM) for factual knowledge.

Language Models as Knowledge Bases?

Given a cloze statement that queries the model for a missing token,
knowledgeable LMs rank ground truth tokens high and other tokens lower



Language Models as Knowledge Bases?

SQuAD Question-Answer Pair

("Who developed the theory of relativity?", Einstein)

Question " The theory of relativity was developed by [MASK] "

Answer Einstein

Original Fact

(Francesco Conti, `born-in`, [**Florence**, ~~Italy~~])

Question " Francesco Conti was born in [MASK] "

Answer Florence

ConceptNet Triple

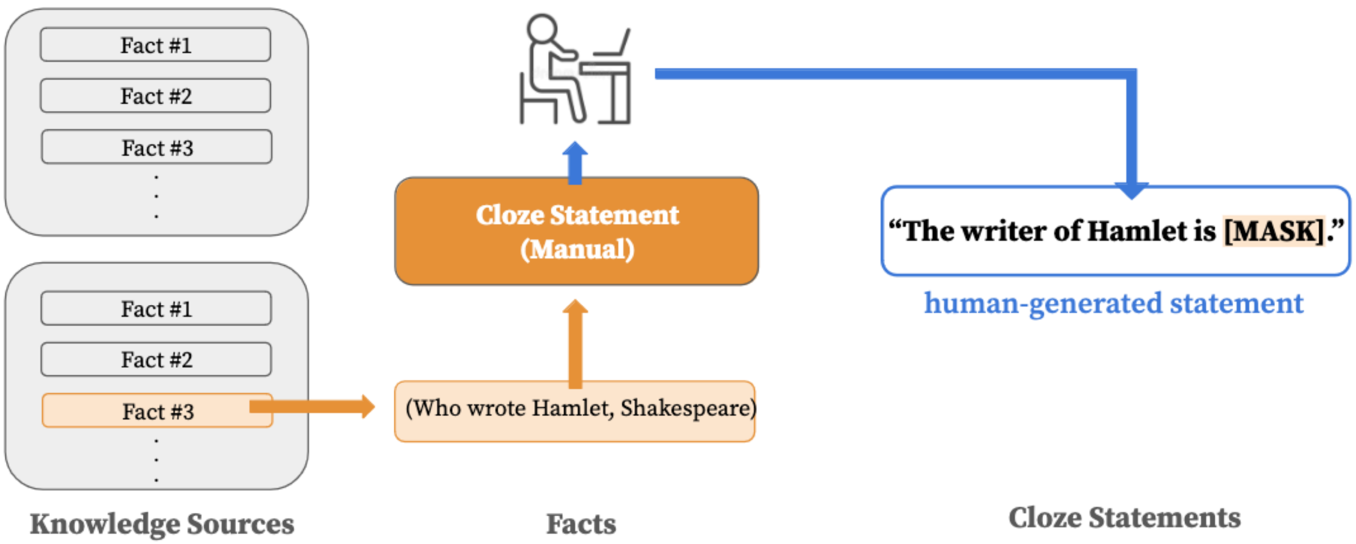
(ravens, `CapableOf`, fly)

Question " Ravens can [MASK] "

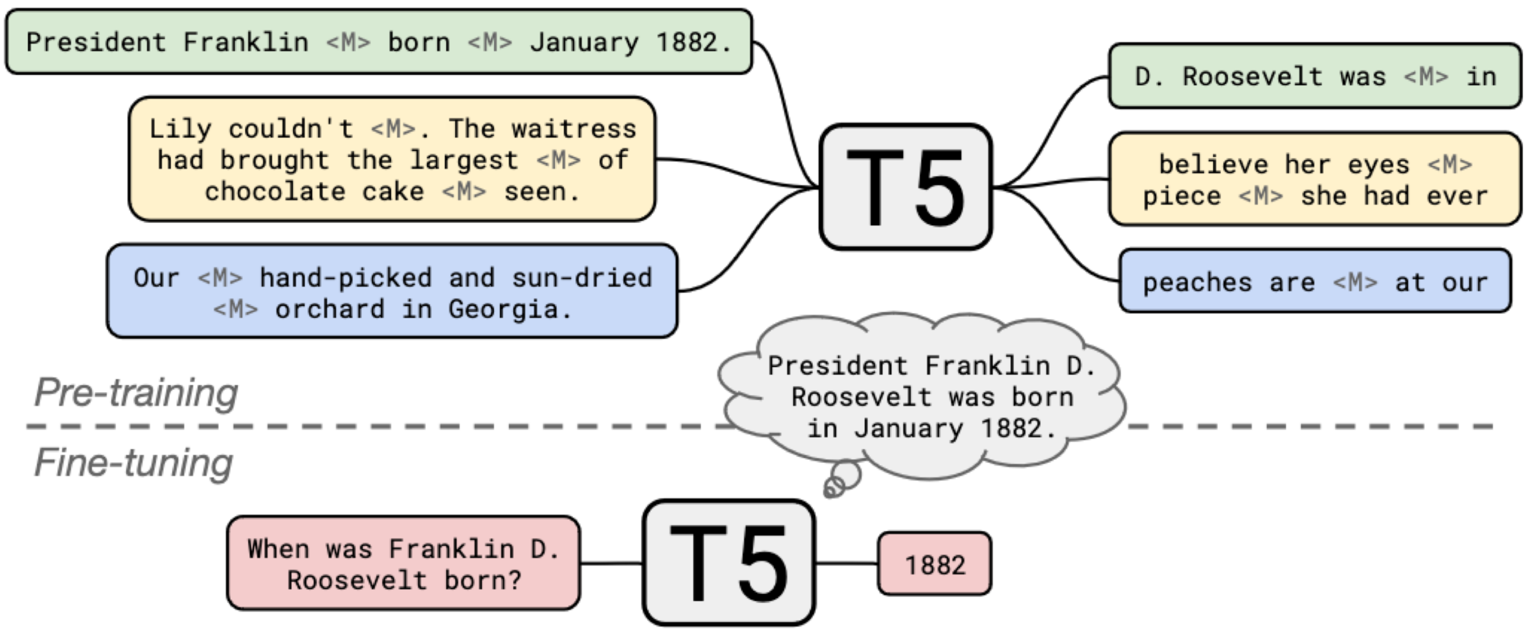
Answer fly

Language Models as Knowledge Bases?

- Knowledge sources: ConceptNet, Google-RE, SQuAD



Language Models as Knowledge Bases?



(Roberts et al., 2020)

Language Models as Knowledge Bases?

- **Pre-training resources**
  - **T5 v1.0:** trained with the unsupervised "span corruption" task on C4 as well as *supervised translation, summarization, classification, and reading comprehension tasks*
  - **T5 v.1.1:** trained only with the C4

- **Model size**
  - Base (220 million parameters)
  - Large (770 million)
  - 3B (3 billion)
  - 11B (11 billion)

- **Additional pre-training**
  - Salient Span Masking (Guu et al. 2020), mask salient spans (named entities & dates)
  - Continue pre-training the T5 for 100k steps

person | location

Henri Hutin invented Brie cheese while living in North of Meuse, France

Comparison with SOTA

Metric: Exact Match

		NQ	WQ	TQA dev	TQA test	
SOTA Retrieval-based Models (can access external documents)	Chen et al. (2017)	–	20.7	–	–	
	Lee et al. (2019)	33.3	36.4	47.1	–	
	Min et al. (2019a)	28.1	–	50.9	–	
	Min et al. (2019b)	31.8	31.6	55.4	–	
	Asai et al. (2019)	32.6	–	–	–	
	Ling et al. (2020)	–	–	35.7	–	
	Guu et al. (2020)	40.4	40.7	–	–	
	Févry et al. (2020)	–	–	43.2	53.4	
	Karpukhin et al. (2020)	**41.5**	42.4	**57.9**	–	
Closed-Book QA models with fine-tuning (relies only on internal parameters)	T5-Base	25.9	27.9	23.8	29.1	non-SSM
	T5-Large	28.5	30.6	28.7	35.9	
	T5-3B	30.4	33.6	35.1	43.4	
	T5-11B	32.6	37.2	42.3	50.1	
	T5-11B + SSM	34.8	40.8	51.0	60.5	SSM
	T5.1.1-Base	25.7	28.2	24.2	30.6	non-SSM
	T5.1.1-Large	27.3	29.5	28.5	37.2	
	T5.1.1-XL	29.5	32.4	36.0	45.1	
	T5.1.1-XXL	32.8	35.6	42.9	52.5	
	T5.1.1-XXL + SSM	35.2	**42.8**	51.9	**61.6**	SSM
Closed-Book QA model without fine-tuning	GPT-3 few-shot	29.9	41.5	**71.2**	-	
SOTA Retrieval-based Models	SOTA	51.4	-	80.1	-	

Scaling Model Size

Metric: Exact Match

	NQ	WQ	TQA dev	TQA test
Chen et al. (2017)	–	20.7	–	–
Lee et al. (2019)	33.3	36.4	47.1	–
Min et al. (2019a)	28.1	–	50.9	–
Min et al. (2019b)	31.8	31.6	55.4	–
Asai et al. (2019)	32.6	–	–	–
Ling et al. (2020)	–	–	35.7	–
Guu et al. (2020)	40.4	40.7	–	–
Févry et al. (2020)	–	–	43.2	53.4
Karpukhin et al. (2020)	**41.5**	42.4	**57.9**	–
T5-Base	25.9	27.9	23.8	29.1
T5-Large	28.5	30.6	28.7	35.9
T5-3B	30.4	33.6	35.1	43.4
T5-11B	32.6	37.2	42.3	50.1
T5-11B + SSM	34.8	40.8	51.0	60.5
T5.1.1-Base	25.7	28.2	24.2	30.6
T5.1.1-Large	27.3	29.5	28.5	37.2
T5.1.1-XL	29.5	32.4	36.0	45.1
T5.1.1-XXL	32.8	35.6	42.9	52.5
T5.1.1-XXL + SSM	35.2	**42.8**	51.9	**61.6**

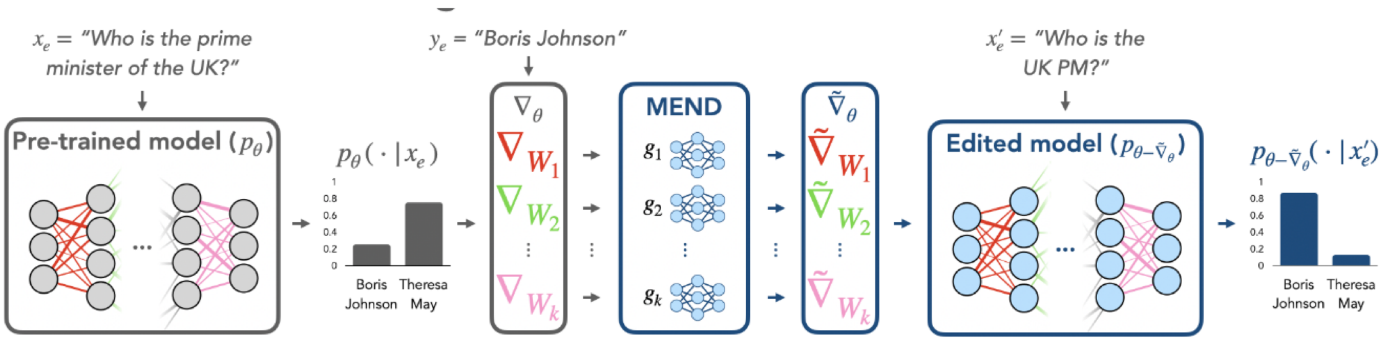
increasing size

increasing performance

increasing size

increasing performance

Knowledge Editing



GPT-3

Who is the president of the United States?

The current president of the United States is Donald Trump.

Who is the president of the United States in 2022?

The current president of the United States is Donald Trump. In 2022, the president will be either Trump or his successor.

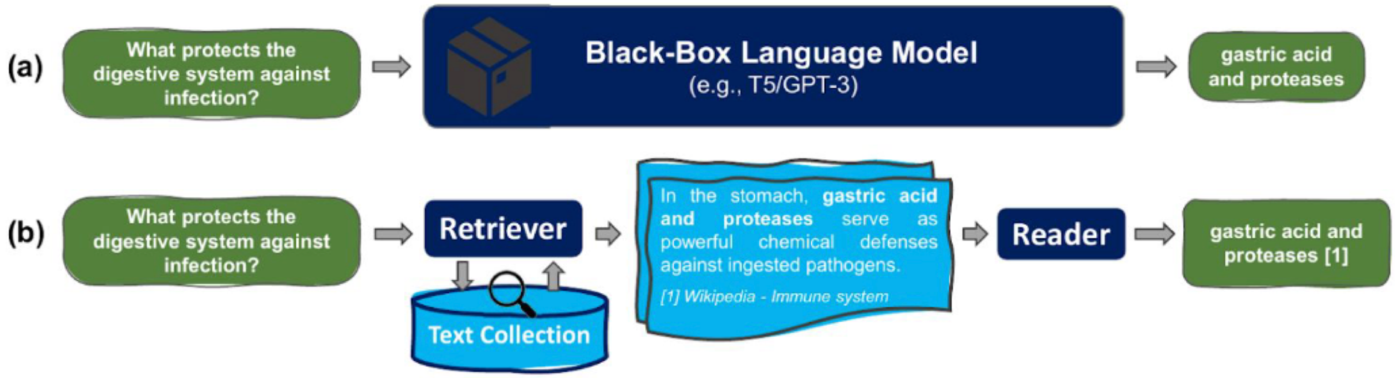
MEND (Mitchell et al, 2022)

Language Models vs. Knowledge Bases

- LMs
  - Scale more easily: removed knowledge extraction
  - More flexible querying: Natural language as query language
- KBs
  - Easier to modify: edit KB directly
  - Easier to trust / understand

Retrieval Augmented Language Models

- Disentangle knowledge with language understanding
- Encode knowledge and explicitly in text
- Retrieve relevant text to generate knowledgeable response
- Easier to update and control
- More efficient



Retrieval Augmented Language Models

	# Retrieval tokens	Granularity	Retriever training	Retrieval integration
Continuous Cache	$\mathcal{O}(10^3)$	Token	Frozen (LSTM)	Add to probs
*k*NN-LM	$\mathcal{O}(10^9)$ Type 1	Token	Frozen (Transformer)	Add to probs
SPALM	$\mathcal{O}(10^9)$	Token	Frozen (Transformer)	Gated logits
DPR	$\mathcal{O}(10^9)$	Prompt	Contrastive proxy	Extractive QA
REALM	$\mathcal{O}(10^9)$	Prompt	End-to-End	Prepend to prompt
RAG	$\mathcal{O}(10^9)$ Type 2	Prompt ?	Fine-tuned DPR	Cross-attention
FiD	$\mathcal{O}(10^9)$	Prompt	Frozen DPR	Cross-attention
EMDR[2]	$\mathcal{O}(10^9)$	Prompt	End-to-End (EM)	Cross-attention
RETRO (ours)	$\mathcal{O}(10^{12})$	**Chunk**	**Frozen (BERT)**	**Chunked cross-attention**

Type 1: *Token*-level Retrieval (mainly) for *LM* – augmenting prediction of next token

Type 2: *Passage*-level Retrieval (mainly) for ***QA*** – retrieving passages relevant to the question

Token-Level Granularity

$$p_{\text{kNN}}(y|x) \propto \sum_{(k_i,v_i)\in\mathcal{N}} \mathbb{1}_{y=v_i} \exp(-d(k_i, f(x)))$$

$$p(y|x) = \lambda\, p_{\text{kNN}}(y|x) + (1-\lambda)\, p_{\text{LM}}(y|x)$$

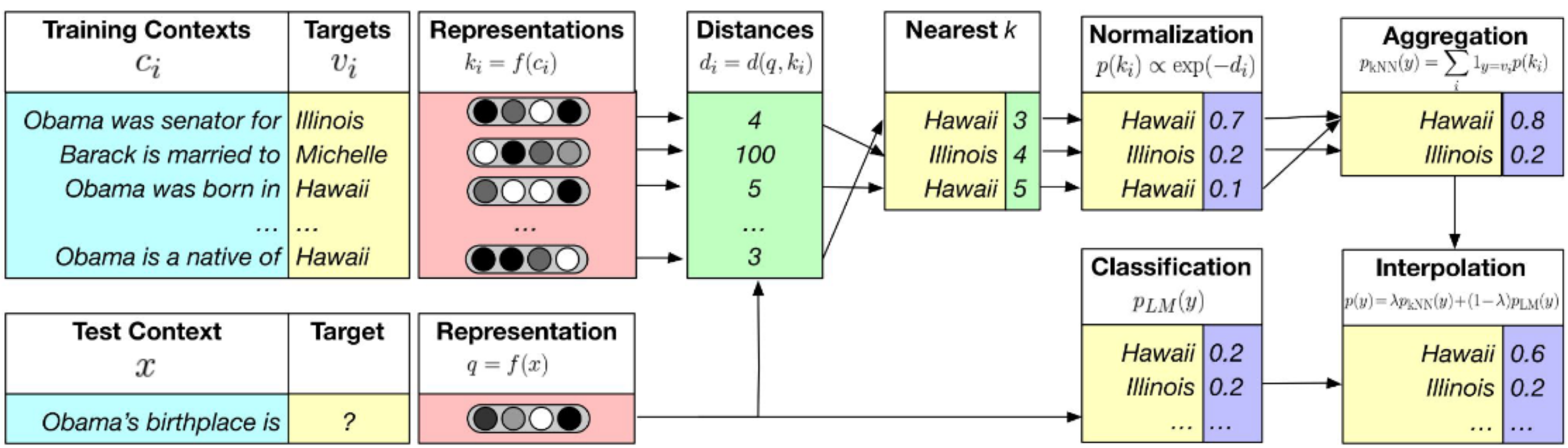
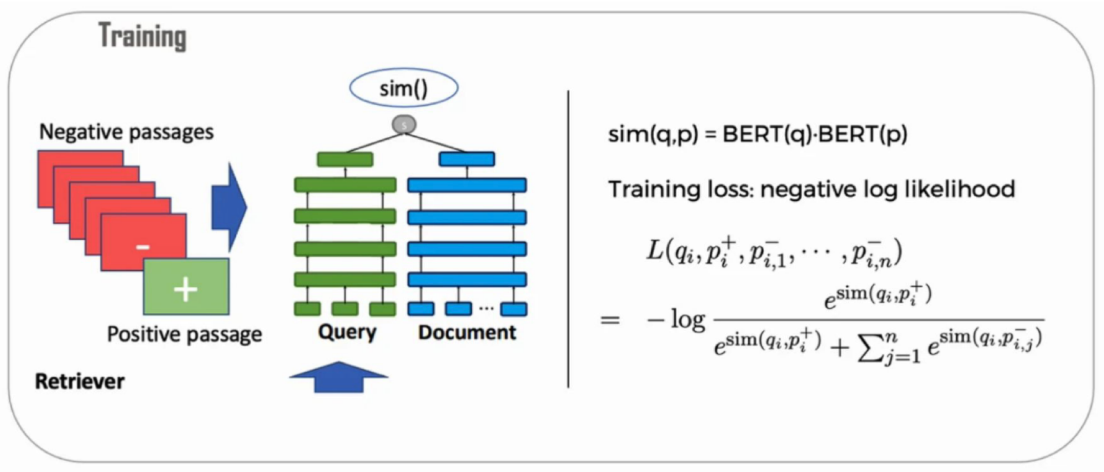


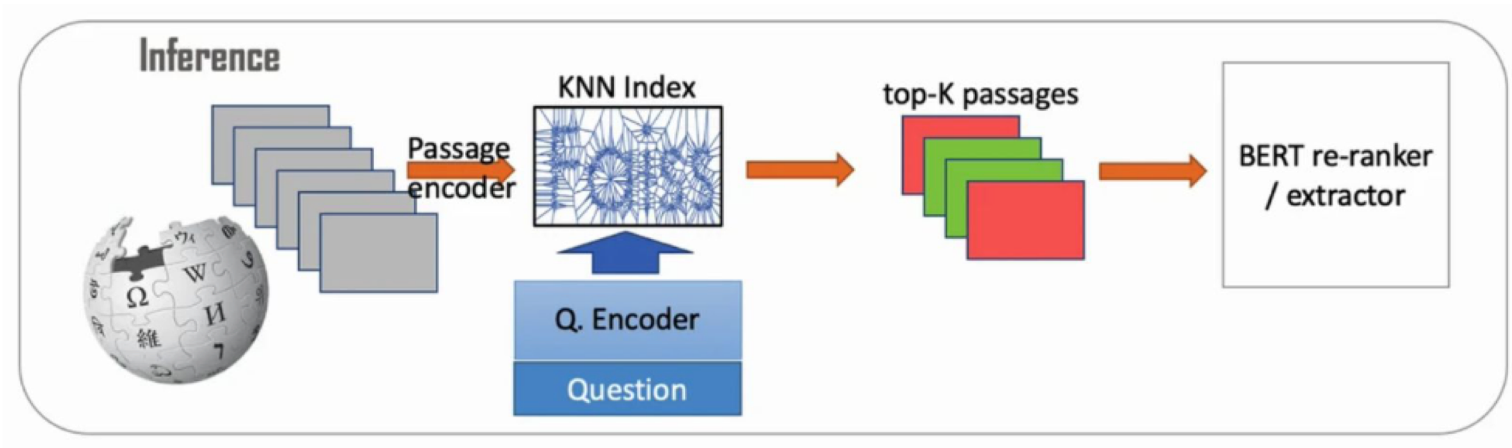
Figure from kNN-LM paper (Khandelwal et al. 2019)

Prompt-Level Granularity



Dense Passage Retriever (Karpukhin et al., 2020)

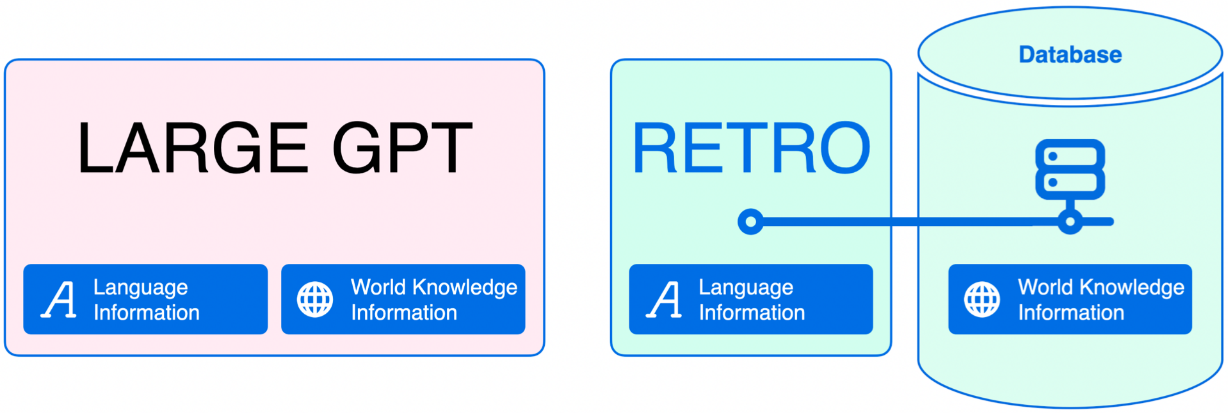
Prompt-Level Granularity



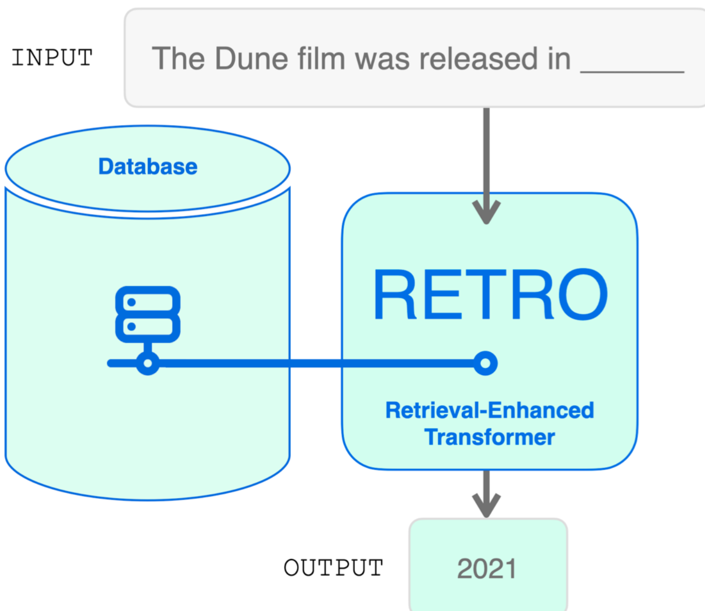
Dense Passage Retriever (Karpukhin et al., 2020)

RETRO

- Improving Language Models By Training From Trillions of Tokens (Borgeaud et al., 2022)



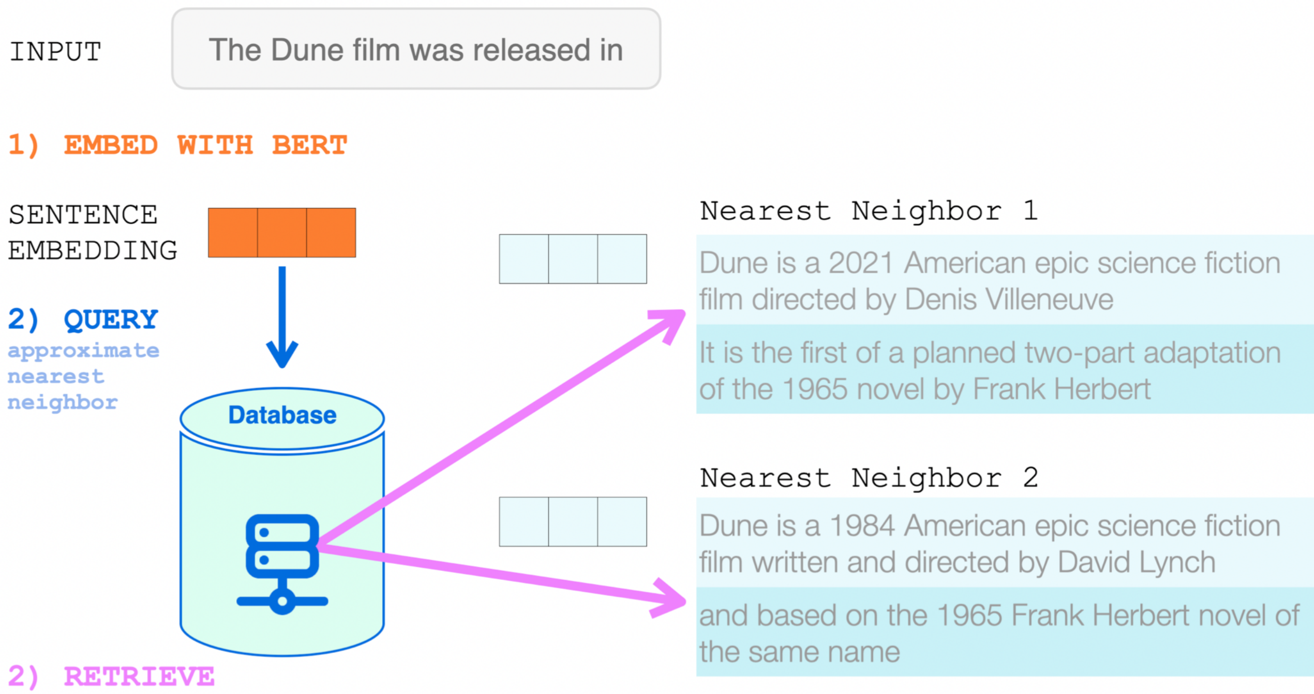
RETRO


INPUT
The Dune film was released in ______
Database
RETRO
Retrieval-Enhanced
Transformer
OUTPUT
2021

RETRO

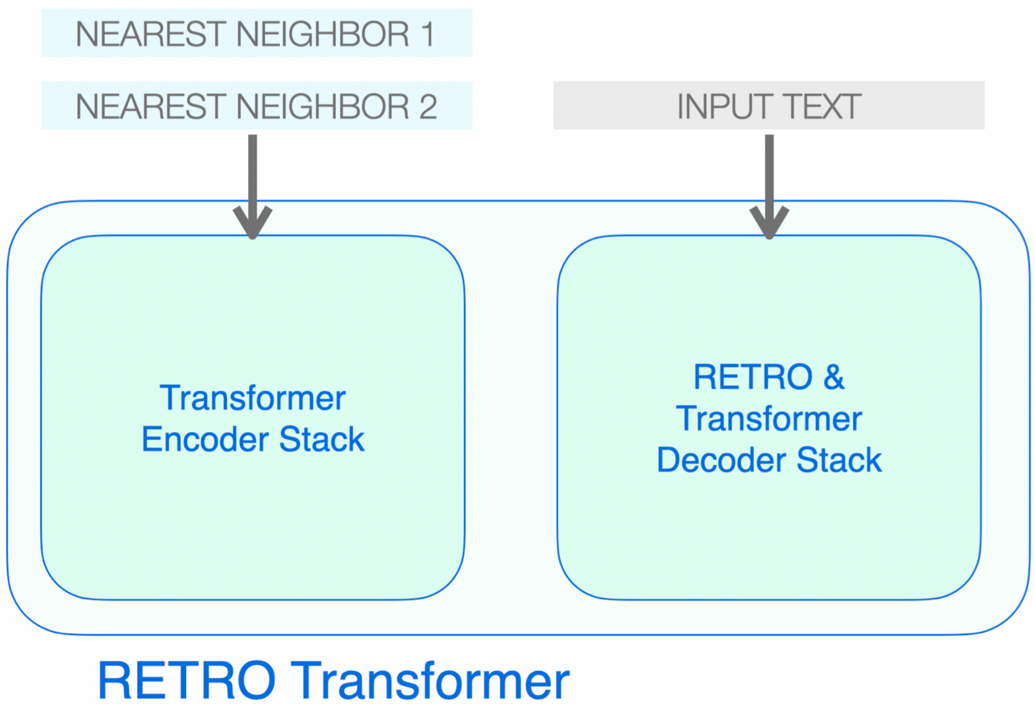
Key (BERT sentence embedding)	Value (text. neighbor and completion chunks. Each up to 64 tokens in length)	
	Dune is a 2021 American epic science fiction film directed by Denis Villeneuve	NEIGHBOR
	It is the first of a planned two-part adaptation of the 1965 novel by Frank Herbert	COMPLETION
	Dune is a 1965 science fiction novel by American author Frank Herbert	NEIGHBOR
	originally published as two separate serials in Analog magazine	COMPLETION
...	...	

RETRO

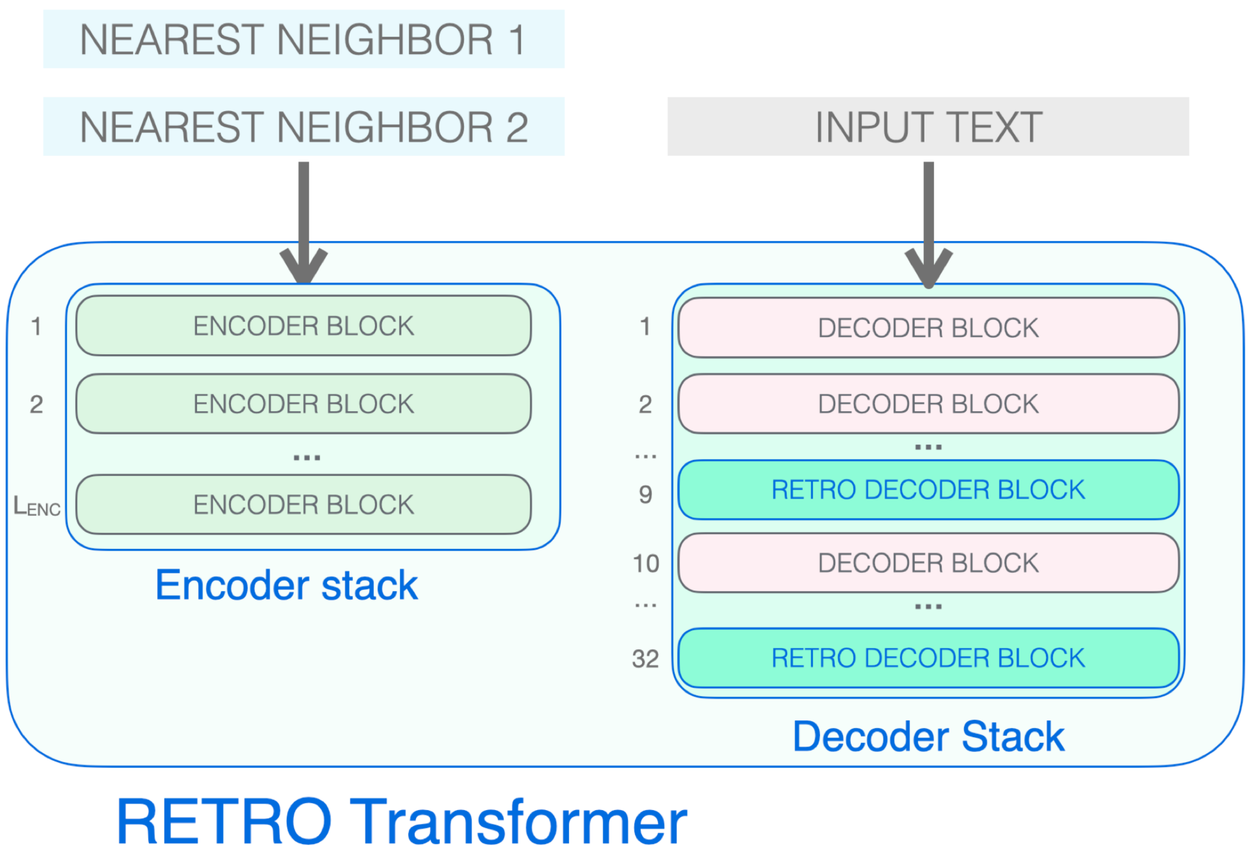


INPUT
The Dune film was released in
1) EMBED WITH BERT
SENTENCE EMBEDDING
2) QUERY
approximate nearest neighbor
Database
2) RETRIEVE
Nearest Neighbor 1
Dune is a 2021 American epic science fiction film directed by Denis Villeneuve
It is the first of a planned two-part adaptation of the 1965 novel by Frank Herbert
Nearest Neighbor 2
Dune is a 1984 American epic science fiction film written and directed by David Lynch
and based on the 1965 Frank Herbert novel of the same name

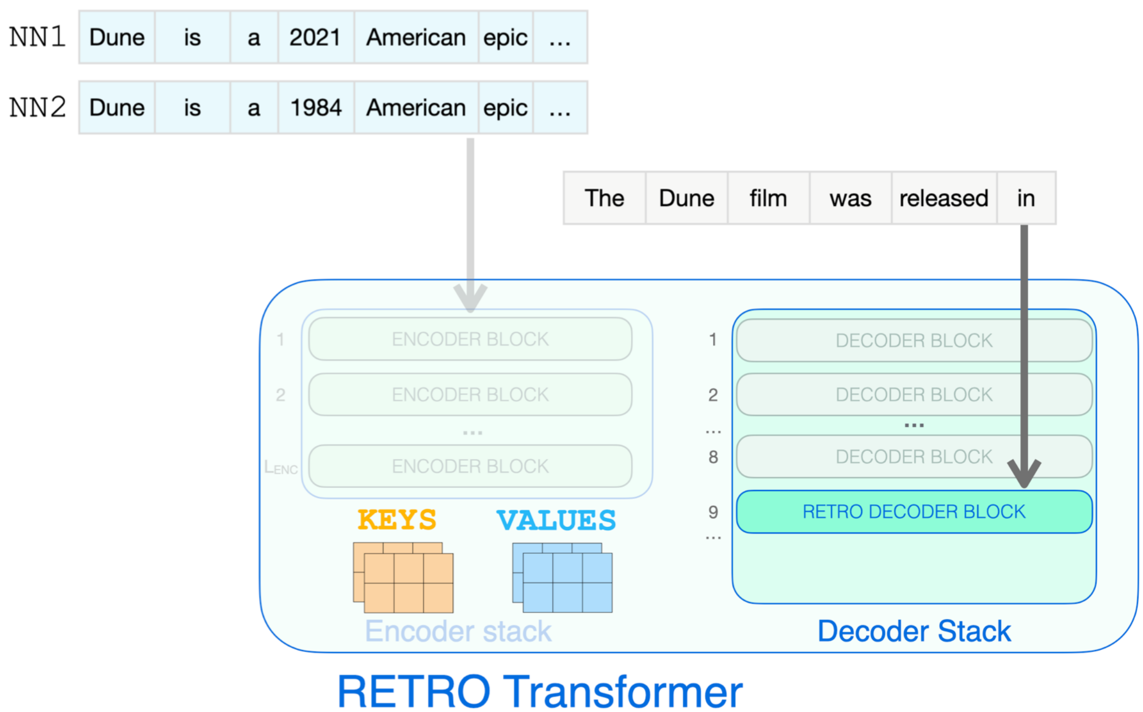
RETRO


NEAREST NEIGHBOR 1
NEAREST NEIGHBOR 2
INPUT TEXT
Transformer Encoder Stack
RETRO & Transformer Decoder Stack
RETRO Transformer

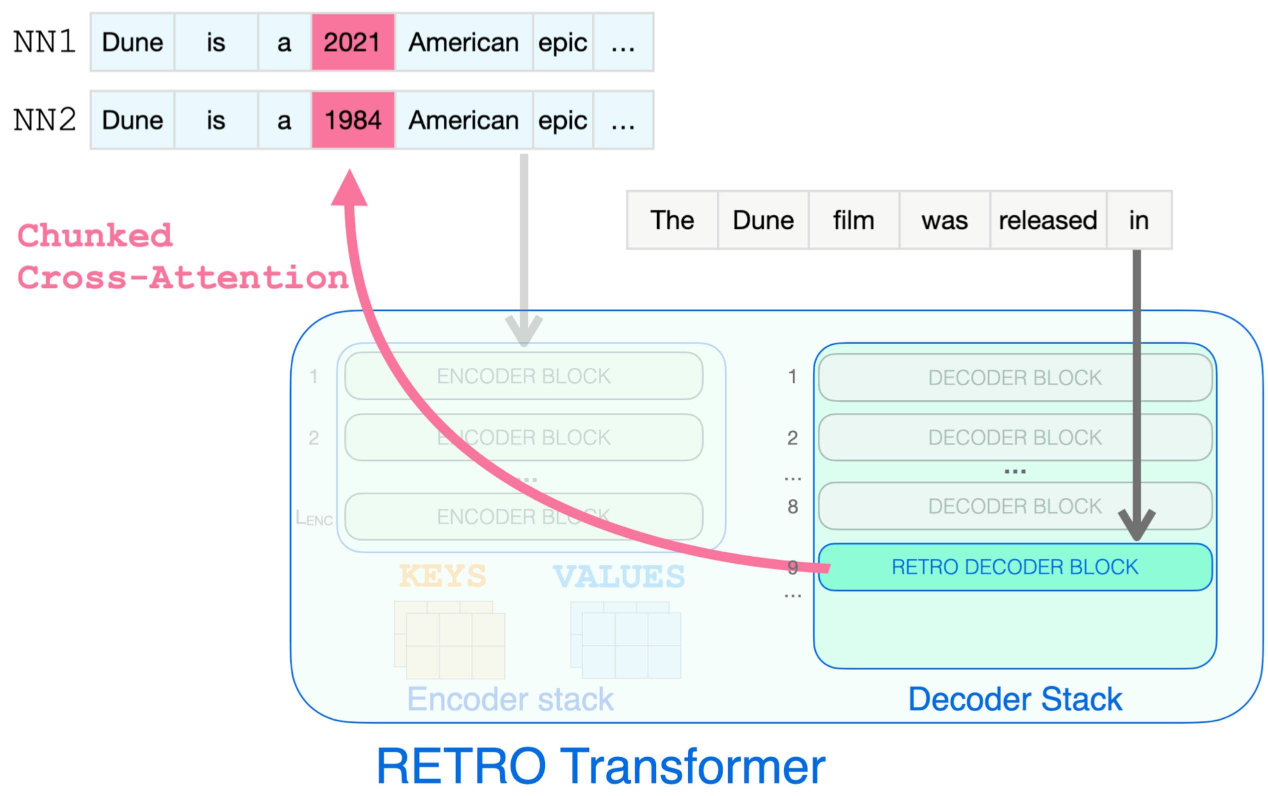
RETRO


NEAREST NEIGHBOR 1
NEAREST NEIGHBOR 2
INPUT TEXT
1
ENCODER BLOCK
2
ENCODER BLOCK
...
LENC
ENCODER BLOCK
Encoder stack
1
DECODER BLOCK
2
DECODER BLOCK
...
9
RETRO DECODER BLOCK
10
DECODER BLOCK
...
32
RETRO DECODER BLOCK
Decoder Stack
RETRO Transformer

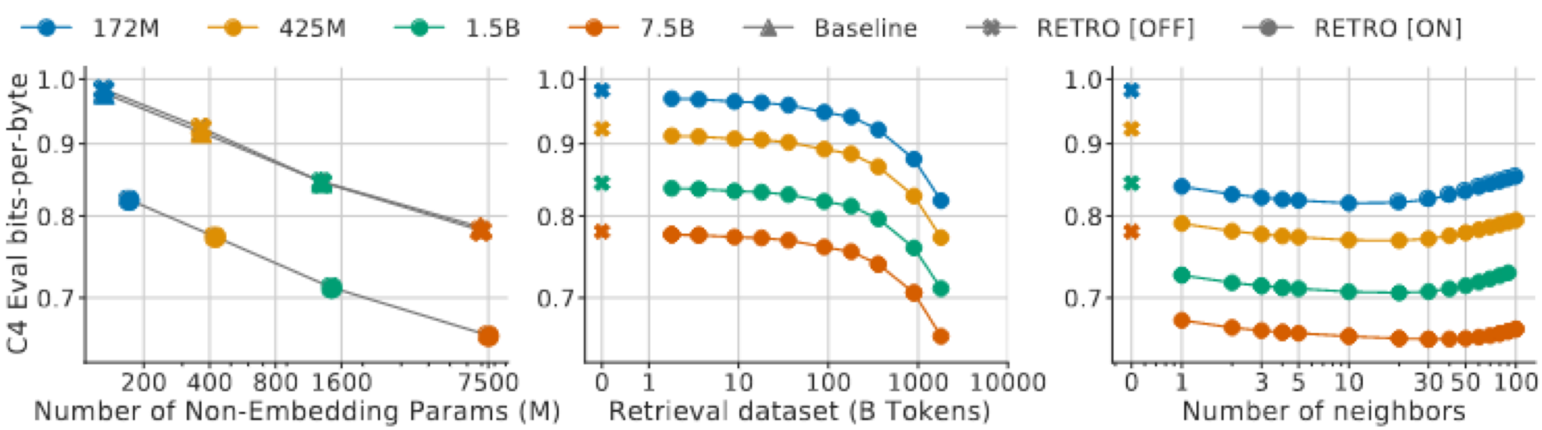
RETRO


NN1 Dune is a 2021 American epic ...
NN2 Dune is a 1984 American epic ...
The Dune film was released in
ENCODER BLOCK
ENCODER BLOCK
ENCODER BLOCK
LENC
KEYS
VALUES
Encoder stack
1 DECODER BLOCK
2 DECODER BLOCK
8 DECODER BLOCK
9 RETRO DECODER BLOCK
Decoder Stack
RETRO Transformer

RETRO

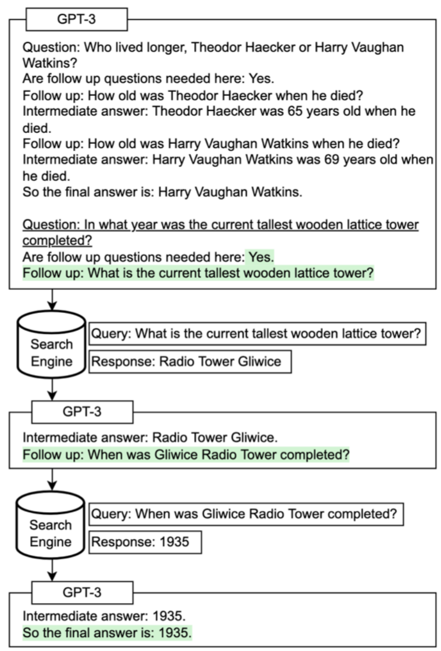

NN1
Dune is a 2021 American epic ...
NN2
Dune is a 1984 American epic ...
Chunked Cross-Attention
The Dune film was released in
ENCODER BLOCK
ENCODER BLOCK
ENCODER BLOCK
KEYS
VALUES
Encoder stack
DECODER BLOCK
DECODER BLOCK
DECODER BLOCK
RETRO DECODER BLOCK
Decoder Stack
RETRO Transformer

RETRO Scaling

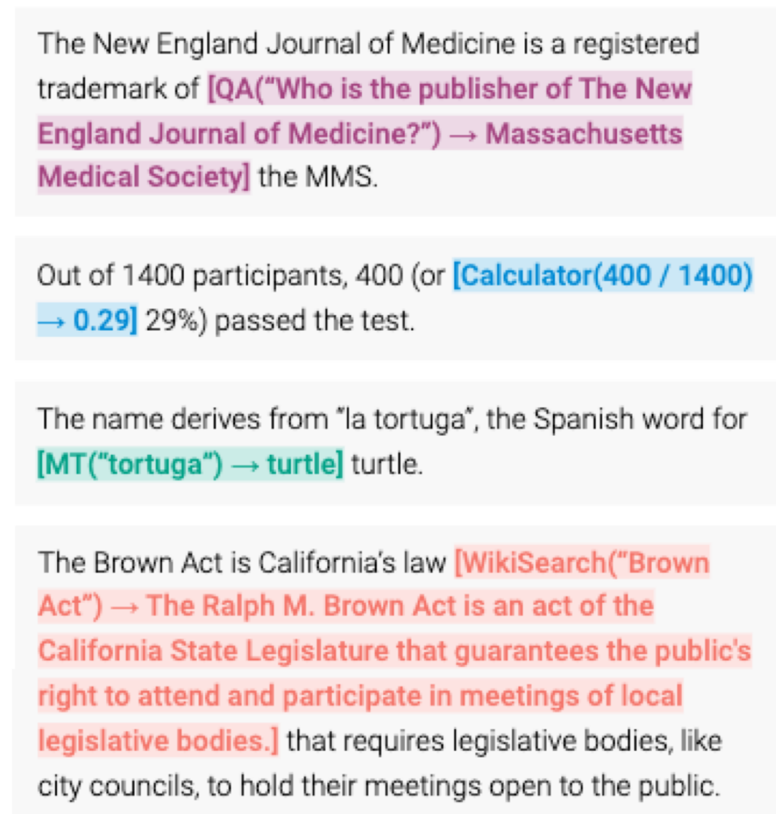


Beyond Augmentation with Text

- Search Engines
- Specialized Models
- Calculators
- Custom Data
- Etc.



(Press et al. 2022)



Toolformer (Schick et al. 2023)

Panel

Slido.com #3854461