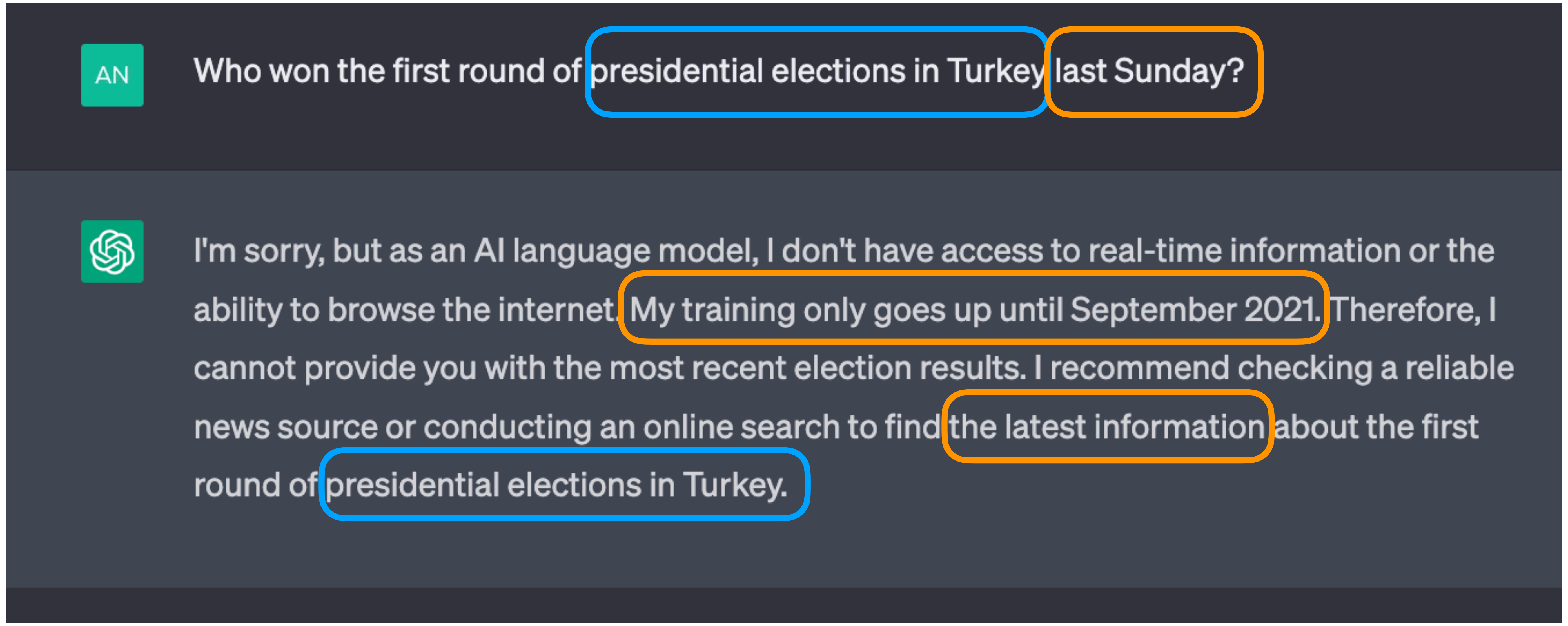
Retrieval-Augmented LMs

Antoine Bosselut

Reading Comprehension challenges

Factuality **Temporality**



What are challenges of answering this question?

- Can we update the model's knowledge without updating its parameters?

Limitations of PLMs (& LLMs)

- **Hallucination** problem (e.g., factual errors)
- Long-tail knowledge (e.g., domain-specific) may not be well-represented in the model's pretraining corpus
- Cannot easily expand or update parameters after pretraining
  - knowledge learned during pretraining is **static**
- Source of information is **non-attributable**

Factual-heavy NLP tasks

Fact Verification

Claim: The Rodney King riots took place in the most populous county in the USA.

[wiki/Los_Angeles_Riots]
The 1992 Los Angeles riots, also known as the Rodney King riots were a series of riots, lootings, arsons, and civil disturbances that occurred in Los Angeles County, California in April and May 1992.

[wiki/Los_Angeles_County]
Los Angeles County, officially the County of Los Angeles, is the most populous county in the USA.

Verdict: Supported

FEVER dataset

Factual Question Answering

Huguenot numbers peaked near an estimated two million by 1562, concentrated mainly in the southern and central parts of France, about one-eighth the number of French Catholics. As Huguenots gained influence and more openly displayed their faith, Catholic hostility grew, in spite of increasingly liberal political concessions and edicts of toleration from the French crown. A series of religious conflicts followed, known as the Wars of Religion, fought intermittently from 1562 to 1598. The wars finally ended with the granting of the Edict of Nantes, which granted the Huguenots substantial religious, political and military autonomy.

Where was France's Huguenot population largely centered?
Ground Truth Answers: the southern and central parts of France | southern and central parts of France, | about one-eighth

SQuaD.v2 dataset

Rajpurkar et al. (2016)
Thorne et al. (2018)

How can we tackle these limitations?

Retrieval

- Precise knowledge access mechanism
- Easy update to known knowledge (update the retrieval knowledge base)
- Neural Retrieval starting to outperform traditional IR

Limitations

- Needs supervision or "heuristics"
- Task specific way to integrate into downstream tasks

Today's Outline

- **Lecture: Retrieval Augmented LMs**
  - **Models:** Model types, training objectives, different external knowledge
  - **Downstream tasks**
  - **Augmented LLMs**: Retrieval in the LLM era
  - **Augmentation benefits:** Modularity, Attribution, Parameter efficiency

Finding the answer in 21M documents

Query → **Documents**

"Where the financial crisis of 2008 started?"

Why can’t we do this with closed-book Extractive & Generative QA models?

We don’t have a context!

But we can get one!
(With adaptations)

Finding the answer in 21M documents

Query → **Documents** → **Retrieve relevant documents**

That might contain the answer

"Where the financial crisis of 2008 started?"

Classical Retrieval: Okapi BM25

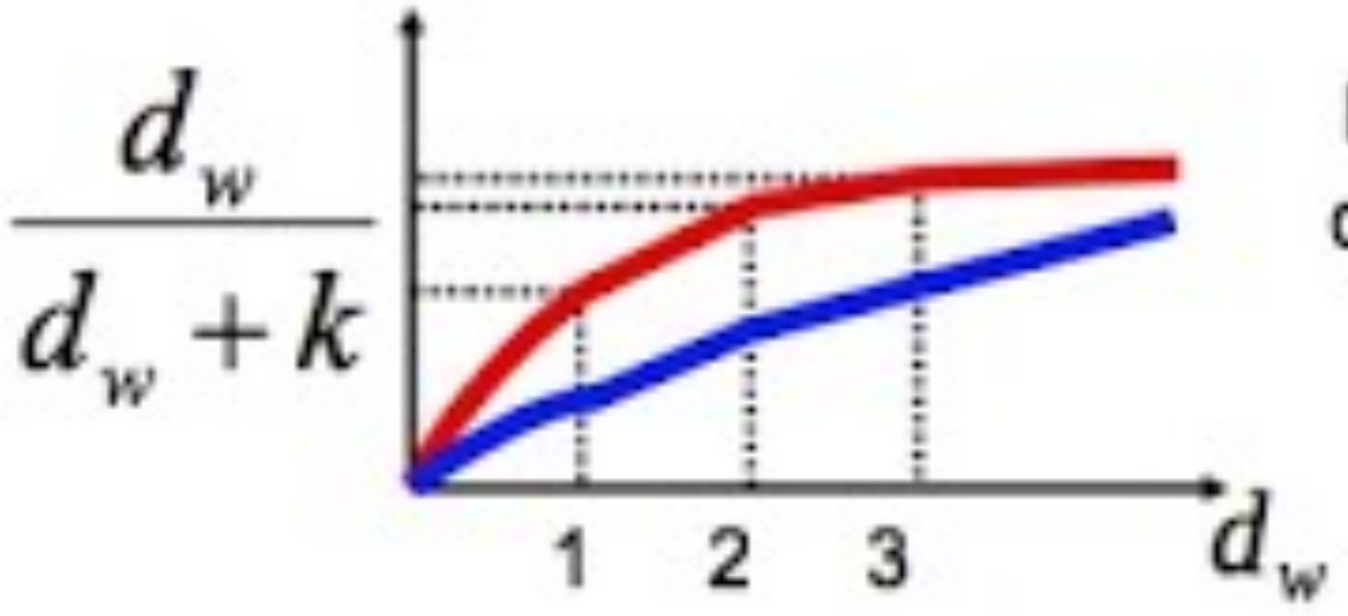
Common words less important

Repetitions of query words ➔ good

$$\log \frac{P(D \mid R=1)}{P(D \mid R=0)} \approx \sum_{w} \left(\underbrace{\frac{d_w(1+k)}{d_w + k((1-b) + \frac{b \cdot dl}{avg.dl})}}_{\text{Repetitions less important than different query words}} \cdot \log \frac{N - N_w + \frac{1}{2}}{N_w + \frac{1}{2}} \right)$$

More words in common with the query ➔ good

But more important if document is relatively long (wrt. average)



BM25 still works quite well for many applications, so don’t ignore classical retrieval if it works better !

Dense Passage Retrieval (DPR)

Documents

Dense Embedding Model

Query

"Where the financial crisis of 2008 started?"

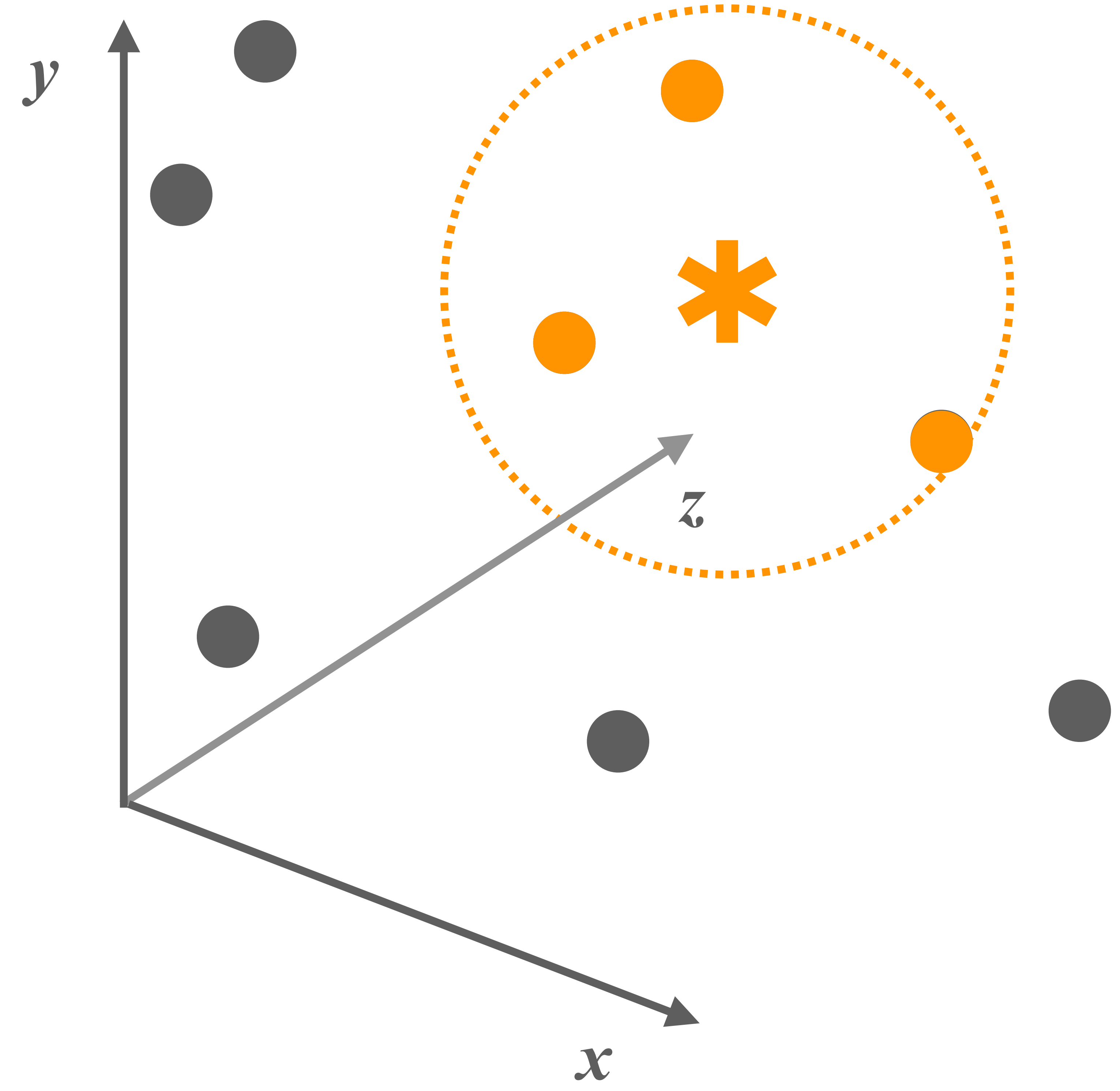
- Create the representations of documents
- Create the representation of the query
- Retrieve *k* documents vectors based on the query vector

Dense Passage Retrieval (DPR)

Documents

$$E_P(\cdot)$$

[-0.5968882 , -0.33086956, -0.32643065, -0.3670732 , ...]



$$\mathrm{sim}(q, p) = E_Q(q)^\intercal E_P(p)$$

Query

$$E_Q(\cdot)$$

[-0.3692328 , -0.37902787, -0.12308089, -0.38124698, ...]

Training DPR

How to create a Document-Query vector space?

<u>Goal:</u> **Relevant** pairs of questions-passages will have a smaller distance compared to the **irrelevant** ones.

"Where the financial crisis of 2008 started?"

Positive passage $p+$

WIKIPEDIA
The Free Encyclopedia

2007–2008 financial crisis

63 languages

Article Talk Tools

From Wikipedia, the free encyclopedia

The **2007–2008 financial crisis**, or **Global Financial Crisis** (GFC), was a severe worldwide economic crisis that occurred in the early 21st century. It was the most serious financial crisis since the Great Depression (1929). Predatory lending targeting low-income homebuyers,[1] excessive risk-taking by global financial institutions,[2] and the bursting of the United States housing bubble culminated in a "perfect storm".

Mortgage-backed securities (MBS) tied to American real estate, as well as a vast web of derivatives linked to those MBS, collapsed in value. Financial institutions worldwide suffered severe damage,[3] reaching a climax with the bankruptcy of Lehman Brothers on September 15, 2008, and a subsequent international banking crisis.[4]

The preconditioning for the financial crisis was complex and multi-causal.[5][6][7] Almost two decades prior, the U.S. Congress had passed legislation encouraging financing for affordable housing.[8] However, in 1999, parts of the Glass-Steagall legislation, which had been adopted in 1933, were repealed, permitting financial institutions to commingle their commercial (risk-averse) and proprietary trading (risk-taking) operations.[9] Arguably the largest contributor to the conditions necessary for financial collapse was the rapid development in predatory financial products which targeted low-income, low-information homebuyers who largely belonged to racial minorities.[10] This market development went unattended by regulators and thus caught the U.S. government by surprise.[11]

Negative passage $p-$

WIKIPEDIA
The Free Encyclopedia

2024 Summer Olympics

76 languages

Article Talk Tools

From Wikipedia, the free encyclopedia

"Paris 2024" and "2024 Olympics" redirect here. For the Summer Paralympics, see 2024 Summer Paralympics. For the Winter Youth Olympics in Gangwon, South Korea, see 2024 Winter Youth Olympics.

The **2024 Summer Olympics** (French: *Jeux olympiques d'été de 2024*), officially the **Games of the XXXIII Olympiad** (French: *Jeux de la XXXIIIe Olympiade*) and commonly known as **Paris 2024**, is an upcoming international multi-sport event that is scheduled to take place from 26 July to 11 August 2024 with Paris as its main host city and 16 cities spread across Metropolitan France and one in Tahiti—an island within the French overseas country and overseas collectivity of French Polynesia—as a subsite.[4]

Paris was awarded the Games at the 131st IOC Session in Lima, Peru, on 13 September 2017. Due to multiple withdrawals that left only Paris and Los Angeles in contention, the International Olympic Committee (IOC) approved a process to concurrently award the 2024 and 2028 Summer Olympics to the two cities. Having previously hosted in 1900 and 1924, Paris will become the second city (after London) to host the Summer Olympics three times. Paris 2024 will mark the centenary of Paris 1924, be the sixth Olympic games hosted by France (three in summer and three in winter), and the first Olympic Games in France since the 1992 Winter Olympics in Albertville. The event will return to its traditional 4-year Olympiad cycle, as the 2020 Summer Olympics in Tokyo was delayed to 2021 due to the COVID-19 pandemic. The Games would be the first of three successive Summer Olympics to use only two official languages (English and French).

DPR LOSS FUNCTION

$$L(q_i, p_i^+, p_{i,1}^-, \cdots, p_{i,n}^-) = -\log \frac{e^{\mathrm{sim}(q_i, p_i^+)}}{e^{\mathrm{sim}(q_i, p_i^+)} + \boxed{\sum_{j=1}^{n} e^{\mathrm{sim}(q_i, p_{i,j}^-)}}}$$

How we can integrate a neural retriever into a Language Model?

Retrieval-Augmented LMs

$$p(y \mid x) =$$

LM

Retriever

Auto-Encoder

$$\sum_{z \in \mathcal{Z}} p(y \mid z, x)$$

$$p(z \mid x)$$

Auto-Encoder

Auto-Regressive

$$\sum_{z \in \mathcal{Z}} \prod_{i}^{N} p(y_i | x, z, y_{1:i-1})$$

Trained to retrieve relevant documents (optional)

z often represented as a latent variable
(may not know what the current document is)

Trained to produce the right answer given the input query and the retrieved documents.

Retrieval-Augmented LMs - Terminology

Information that is stored in the parameters of the models used (both for the LM and the retrieval parts).

knowledge

memory

LM		Retriever
Implicit	vs	Explicit
Parametric	vs	Non-parametric

The type of external source the retriever will use.

modalities

KB

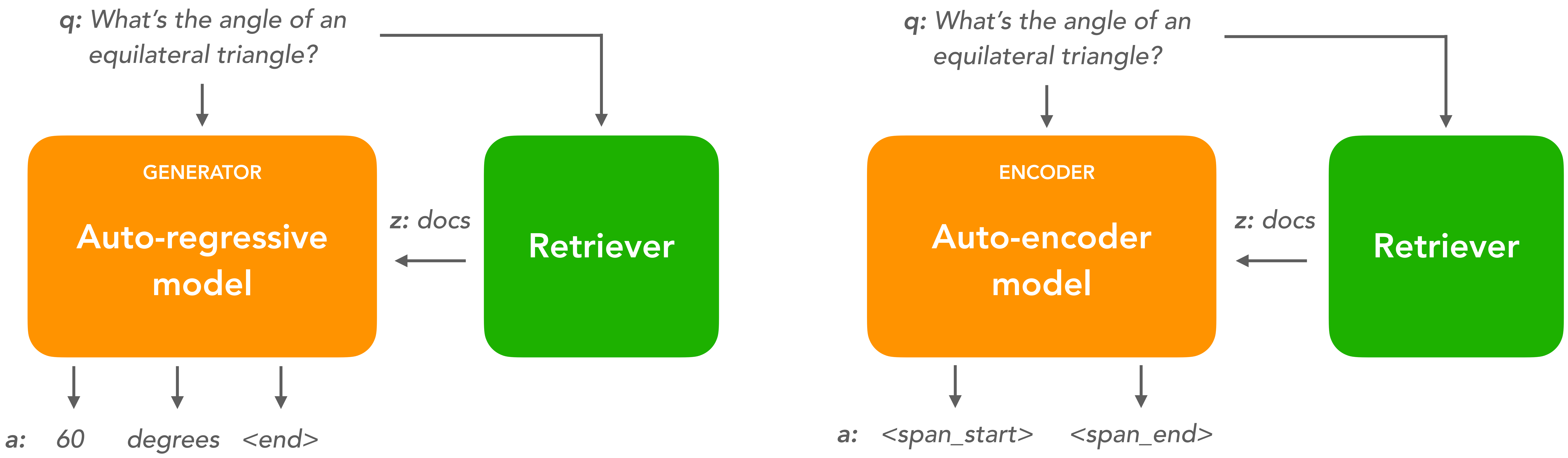
KG

Tools

The landscape of Retrieval-Augmented LMs

ARCHITECTURE OF THE LM	TRAINING OF THE COMPONENTS	TYPES OF EXTERNAL KNOWLEDGE
Generative **vs** **Extractive**	**Pre-training** **vs** **Fine-tuning**	**Document Repositories** **Knowledge Graphs**
RAG: Fine-tuning & KB	*REALM*: Pre-training & KB	*ERNIE*: Pre-training & KG

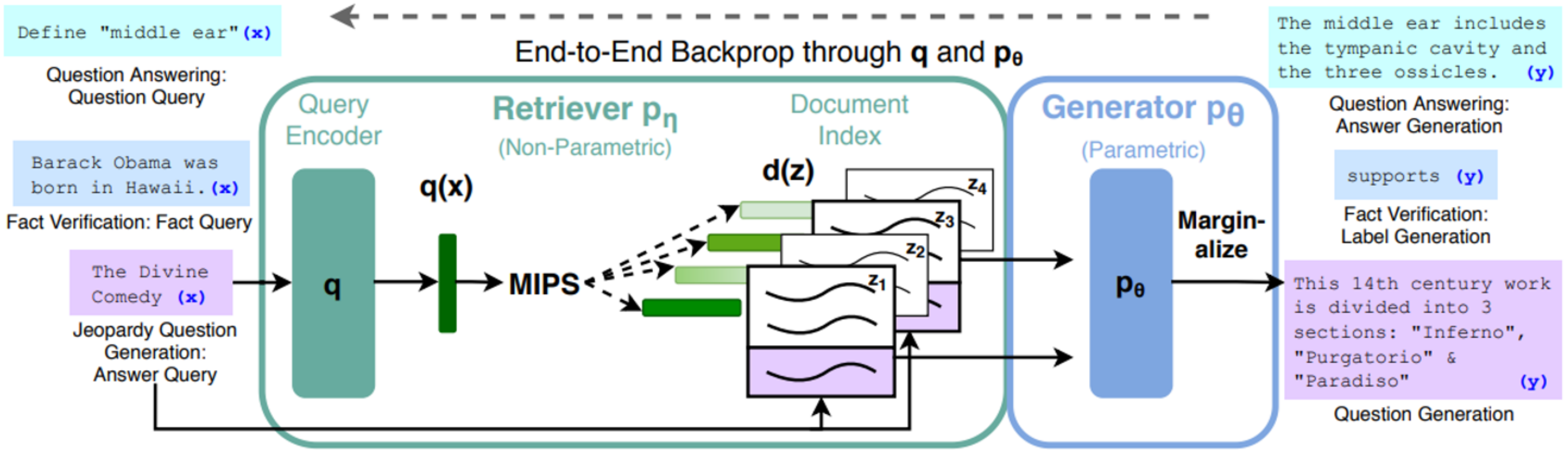
Generative vs Extractive



RAG: Generative Retrieval-Augmented LM

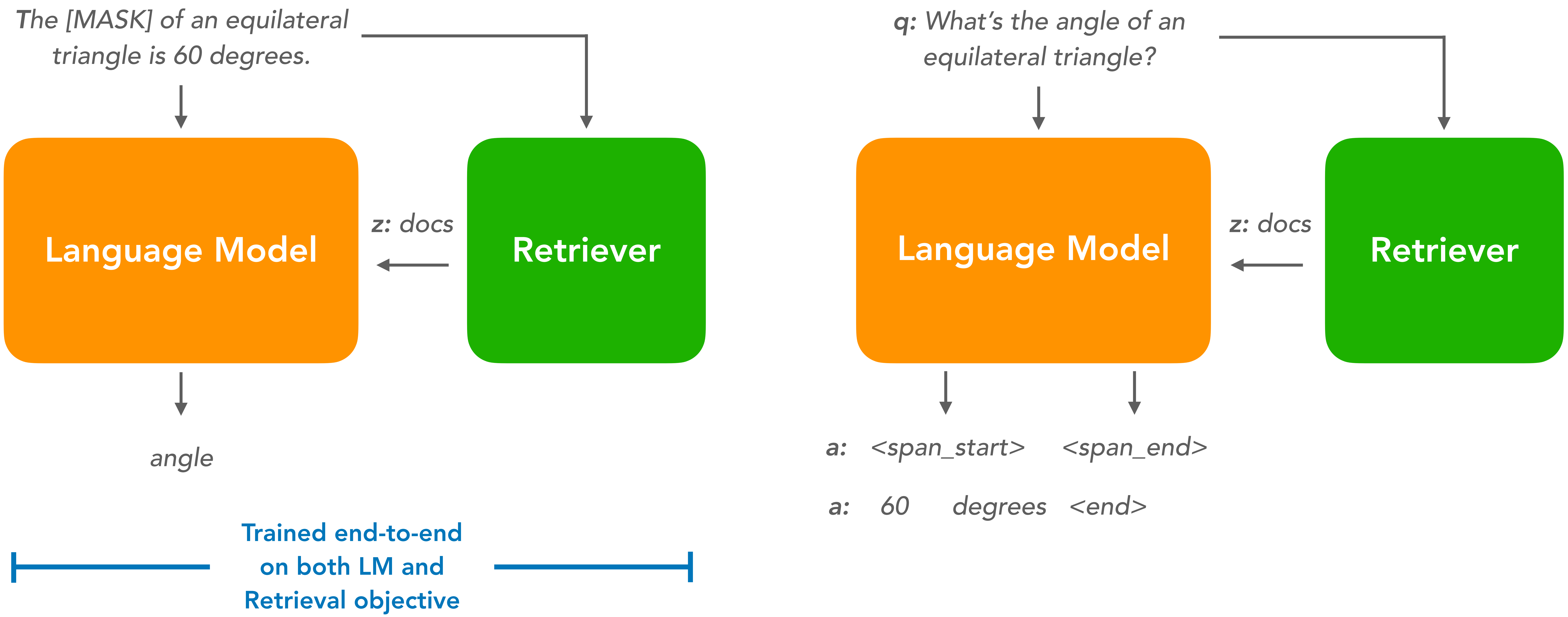
1. Pre-trained generator (e.g. BART)
2. Pre-trained retriever (e.g. DPR)
3. Indexed KB of text documents (e.g. Wikipedia)

$$p(y \mid x) = \sum_{z \in \mathcal{Z}} \prod_{i}^{N} p(y_i | x, z, y_{1:i-1})$$



Lewis et al. (2020)

Pre-training vs Fine-tuning



REALM: Pre-training Retrieval Augmented LMs

First Retrieve:

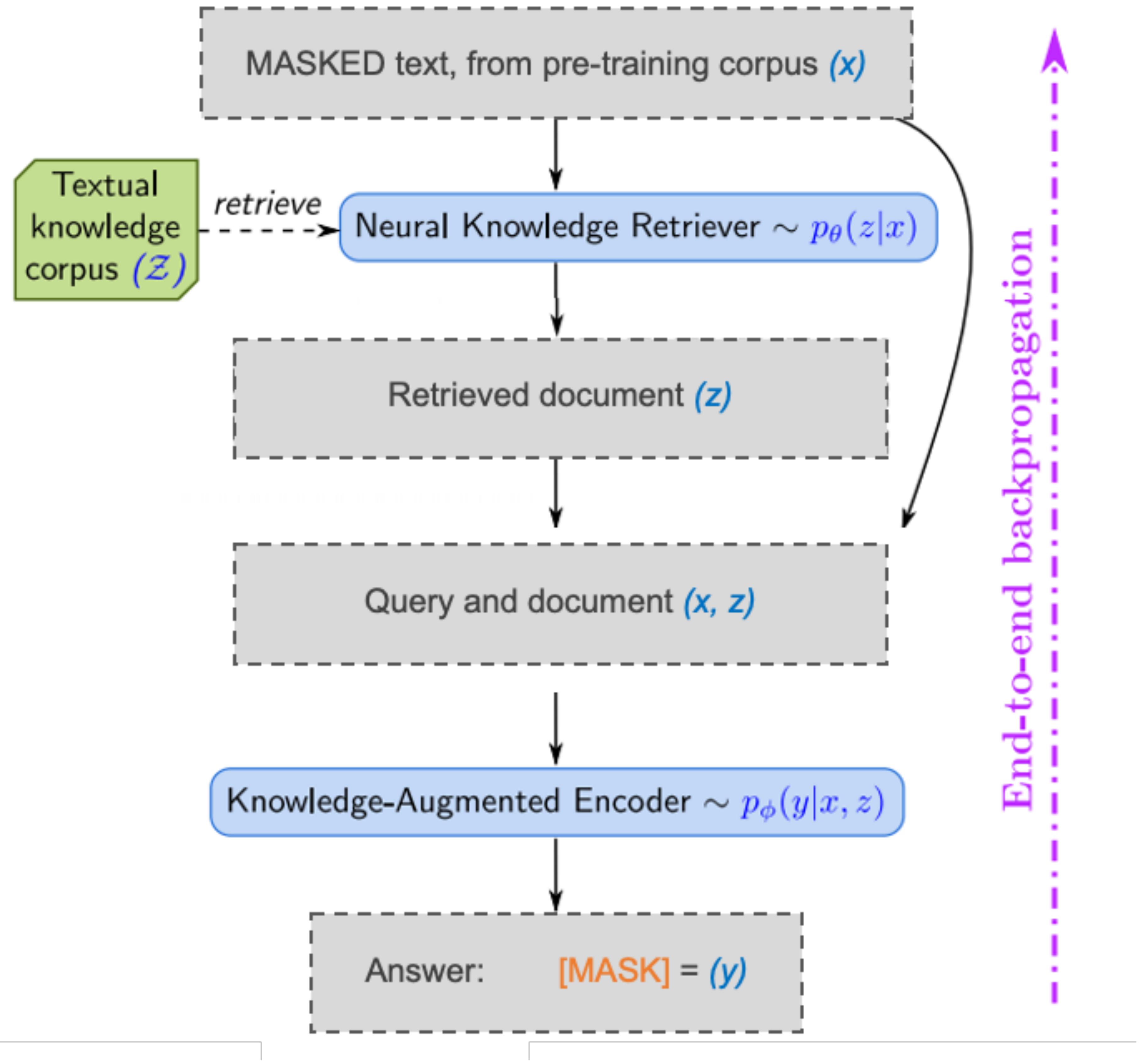
The retriever model is trained on what documents are relevant.
Goal: Penalise uninformative retrievals

Then Predict:

The encoder model is trained to predict the original value of each masked token by attending to the input query and the retrieved documents.
Goal: Minimise perplexity

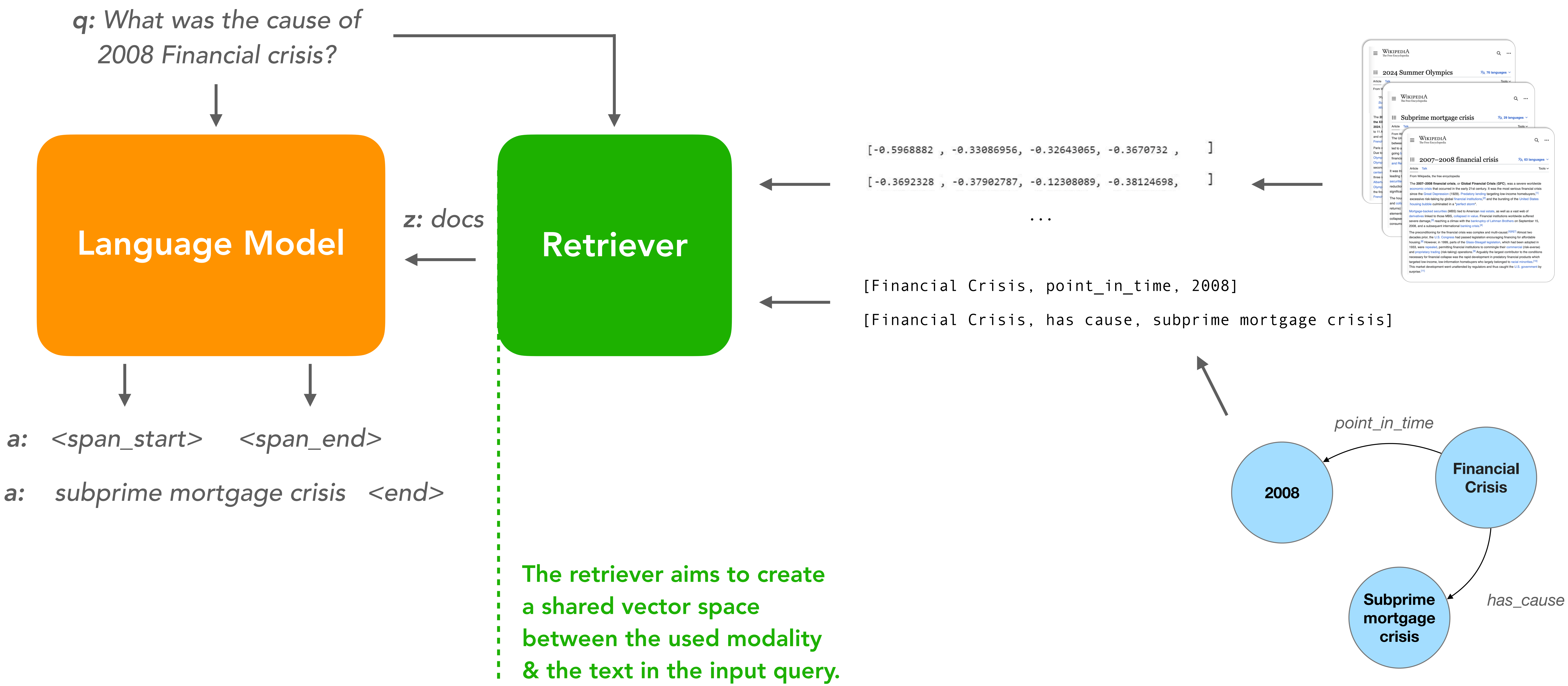
Benefits of pre-training end-to-end

- Transferability across tasks
- Rely on information beyond lexical overlap: the model learns for itself which texts are most useful for reducing perplexity.
- Model-centric **unsupervised alignments** between text in the pre-training corpus X and knowledge corpus Z.



Guu et al. (2020)

Different types of external knowledge

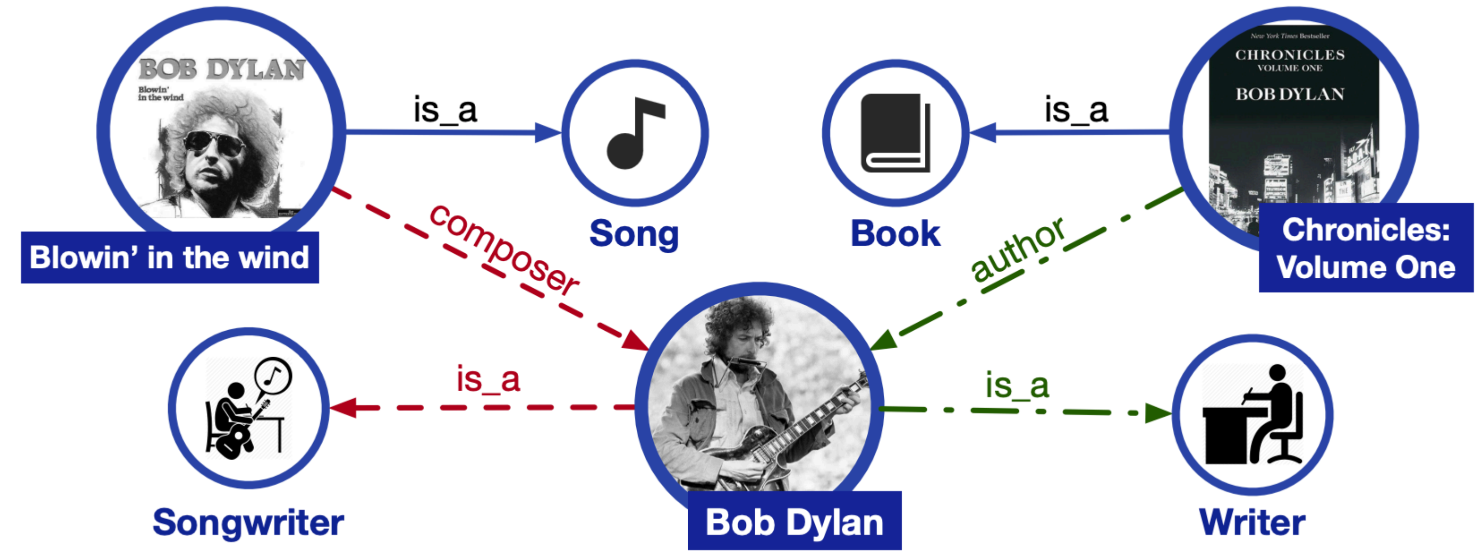


What other modalities could we use as a base ?

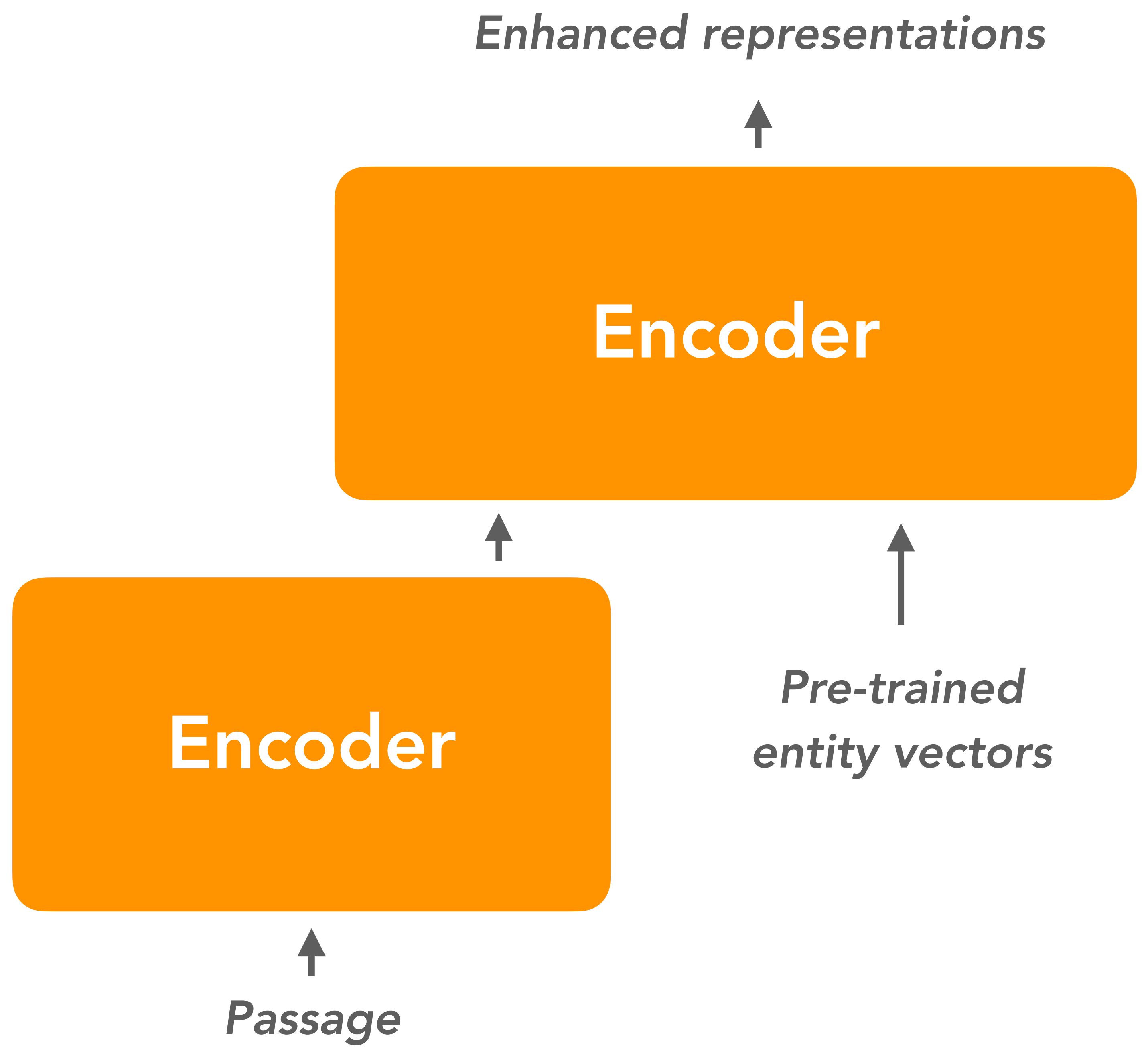
How would we integrate these modalities ?

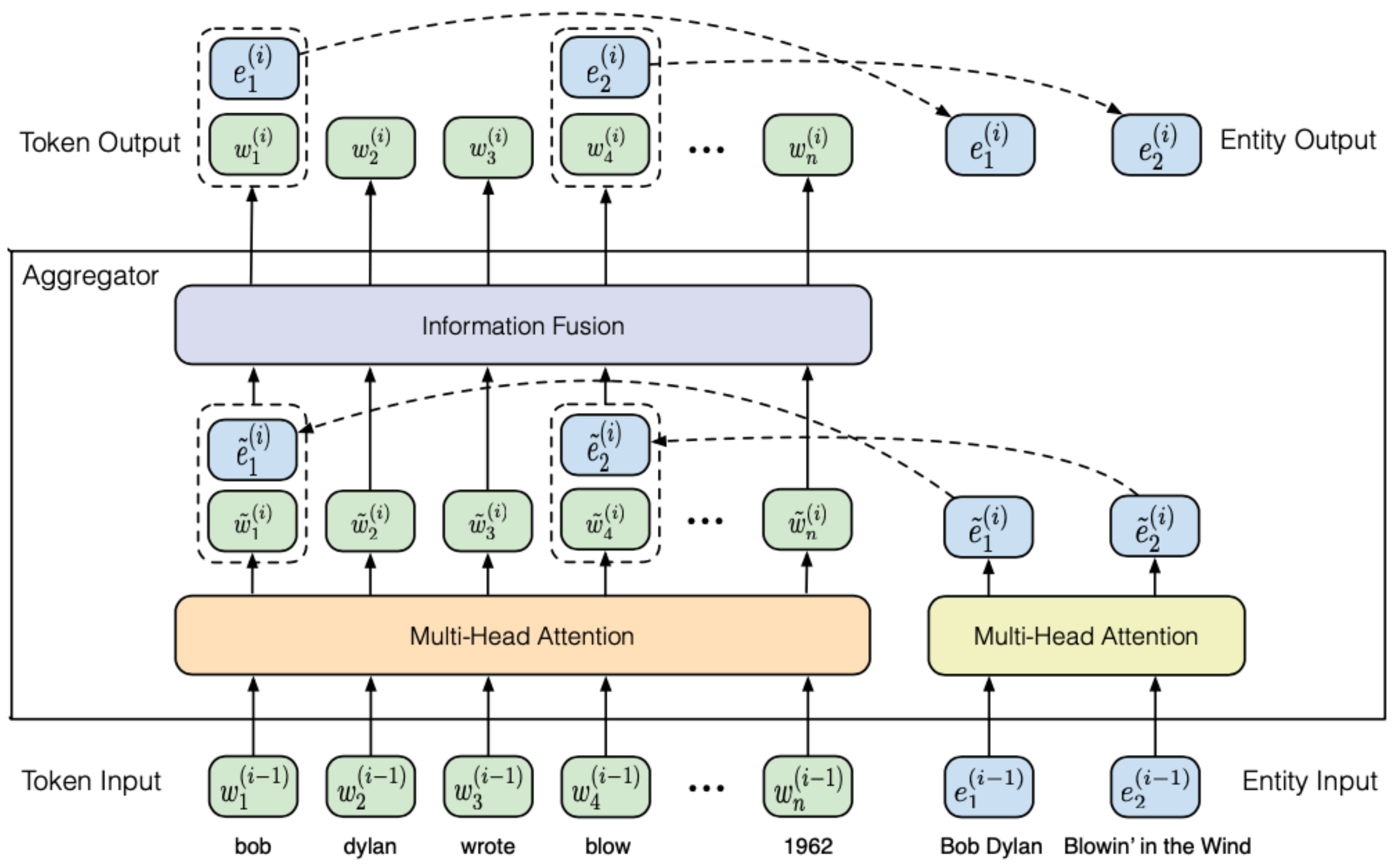
ERNIE: Infuse KG knowledge

1. Extracts the named entity mentions in the text
2. Aligns these mentions to their corresponding entities in KGs.
3. Gets the graph pre-trained entity embeddings for each named entity.
4. Integrates the entity representations in the Encoder model.



Bob Dylan wrote ***Blowin' in the Wind*** in 1962, and wrote ***Chronicles: Volume One*** in 2004.





Bob Dylan wrote ***Blowin' in the Wind*** in 1962

Zhang et al. (2019)

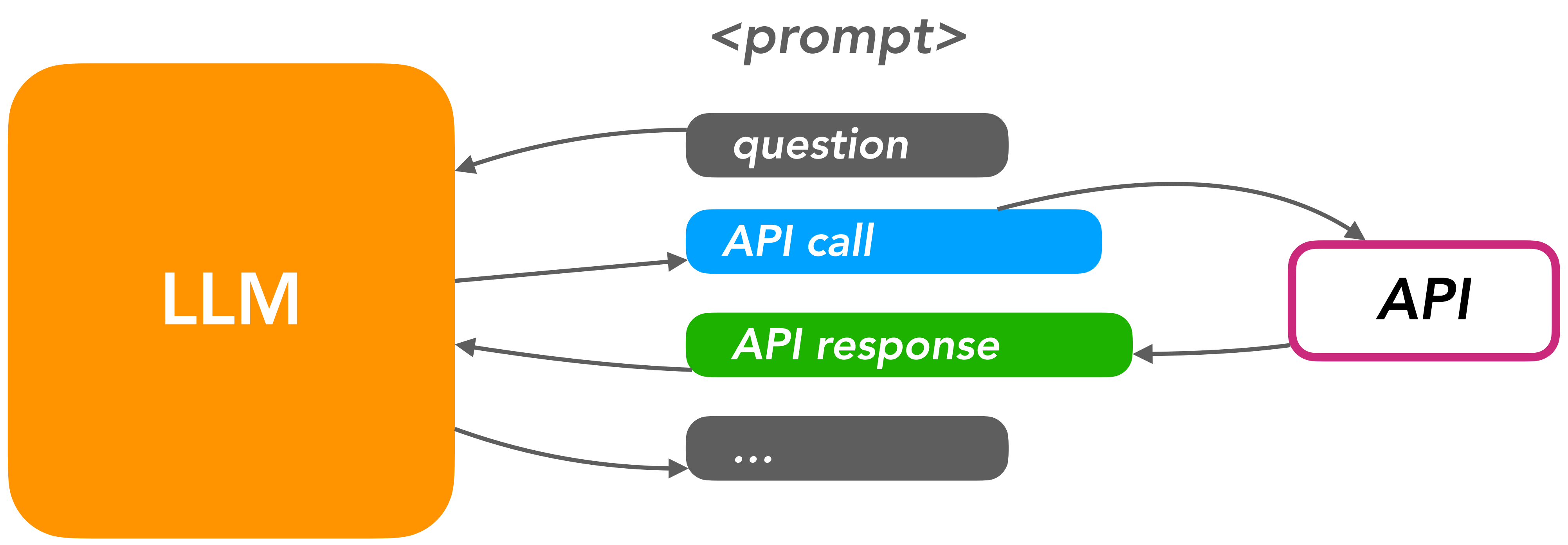
We looked at retrieving information from document bases and knowledge graphs.

What else can be retrieved by a model ?

Augmented LLMs

Retrieve from tools & APIs

Equip language models with the ability to use different tools by means of API calls



Retrieval-Augmented Prompts

```
Act
Question        What is the elevation range for the area that the eastern sector of the
                Colorado orogeny extends into?
Action 1        Search[Colorado orogeny]
Observation 1   The Colorado orogeny was an episode of mountain building (an orogeny) in
                Colorado and surrounding areas.
Action 2        Lookup[eastern sector]
Observation 2   (Result 1 / 1) The eastern sector extends into the High Plains and is called
                the Central Plains orogeny.
Action 3        Search[High Plains]
Observation 3   High Plains refers to one of two distinct land regions:
Action 4        Search[High Plains (United States)]
Observation 4   The High Plains are a subregion of the Great Plains.  From east to west, the
                High Plains rise in elevation from around 1,800 to 7,000 ft (550 to 2,130
                m).[3]
Action 5        Finish[1,800 to 7,000 ft]
```

ReAct (Yao et al. 2024)

The New England Journal of Medicine is a registered trademark of [QA("Who is the publisher of The New England Journal of Medicine?") → Massachusetts Medical Society] the MMS.

Out of 1400 participants, 400 (or [Calculator(400 / 1400) → 0.29] 29%) passed the test.

The name derives from "la tortuga", the Spanish word for [MT("tortuga") → turtle] turtle.

The Brown Act is California's law [WikiSearch("Brown Act") → The Ralph M. Brown Act is an act of the California State Legislature that guarantees the public's right to attend and participate in meetings of local legislative bodies.] that requires legislative bodies, like city councils, to hold their meetings open to the public.

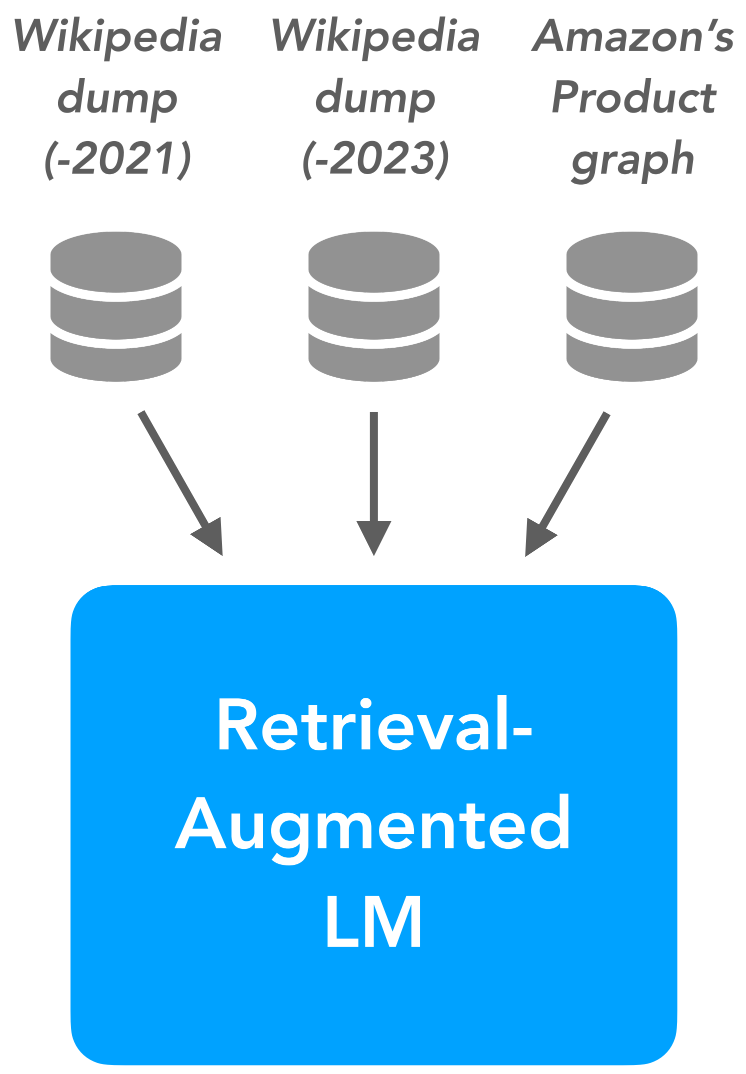
Toolformer (Schick et al. 2023)

What are some benefits of augmentation?

Additional benefits of Augmented LMs

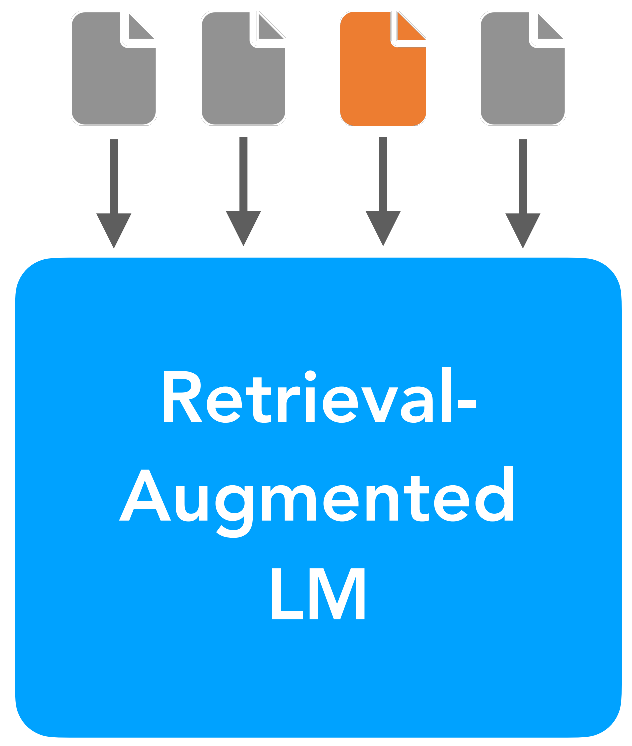
Modularity

We can change external memory and update the model's knowledge on test time.



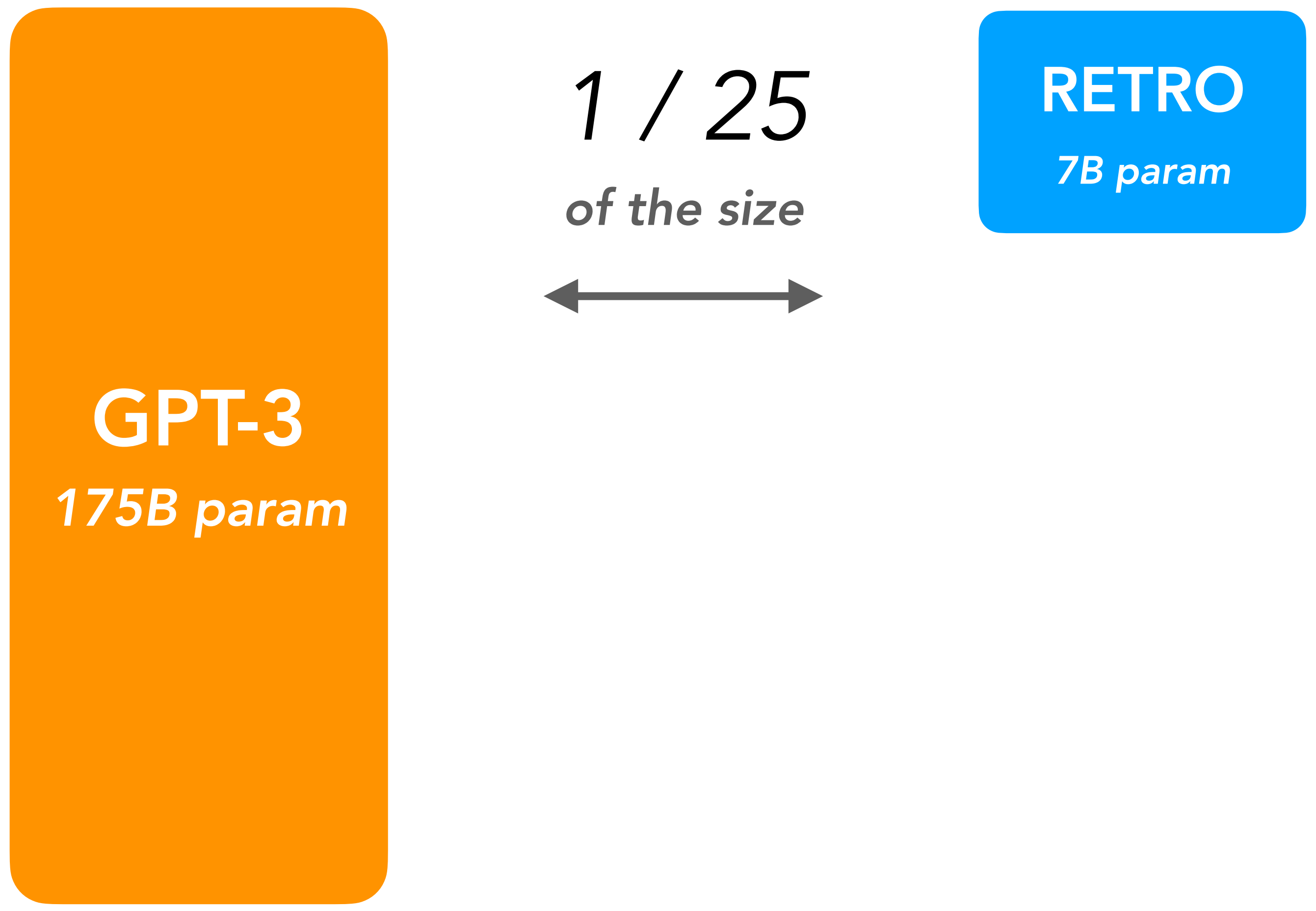
Attribution

We can trace back the information (documents) that the generated answer is based on.



Parameter efficiency

We can leverage external memory to reduce the number of implicit parameters of the LM without compromising performance.



Closed-book vs. Retrieval

Rank	Model	EM	Paper	Code	Result	Year	Tags
1	**Atlas** (full, Wiki-dec-2018 index)	64.0	Atlas: Few-shot Learning with Retrieval Augmented Language Models			2022	
2	**Atlas** (full, Wiki-dec-2021+CC index)	60.4	Atlas: Few-shot Learning with Retrieval Augmented Language Models			2022	
3	**FiE**	58.4	FiE: Building a Global Probability Space by Leveraging Early Fusion in Encoder for Open-Domain Question Answering			2022	
4	**R2-D2** (full)	55.9	R2-D2: A Modular Baseline for Open-Domain Question Answering			2021	
5	**ReAtt**	54.7	Retrieval as Attention: End-to-end Learning of Retrieval and Reading within a Single Transformer			2022	
6	**FiD-KD** (full)	54.7	Leveraging Passage Retrieval with Generative Models for Open Domain Question Answering			2020	
7	**EMDR^2**	52.5	End-to-End Training of Multi-Document Reader and Retriever for Open-Domain Question Answering			2021	
8	**FID** (full)	51.4	Leveraging Passage Retrieval with Generative Models for Open Domain Question Answering			2020	
9	**RETRO + DPR** (full)	45.5	Improving language models by retrieving from trillions of tokens			2021	
10	**code-davinci-002 175B + REPLUG LSR** (few-shot)	45.5	REPLUG: Retrieval-Augmented Black-Box Language Models			2023	
11	**Atlas** (few-shot, k=64, Wiki-Dec-2018 index)	45.1	Atlas: Few-shot Learning with Retrieval Augmented Language Models			2022	few-shot
12	**code-davinci-002 175B + REPLUG** (few-shot)	44.7	REPLUG: Retrieval-Augmented Black-Box Language Models			2023	
13	**RAG**	44.5	Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks			2020	
14	**Blended RAG**	42.63	Blended RAG: Improving RAG (Retriever-Augmented Generation) Accuracy with Semantic Search and Hybrid Query-Based Retrievers			2024	
15	**Atlas** (few-shot, k=64, Wiki-dec-2021+CC index)	42.4	Atlas: Few-shot Learning with Retrieval Augmented Language Models			2022	few-shot
16	**DPR**	41.5	Dense Passage Retrieval for Open-Domain Question Answering			2020	
17	**REALM**	40.4	REALM: Retrieval-Augmented Language Model Pre-Training			2020	
18	**LLaMA 65B** (few-shot, k=64)	39.9	LLaMA: Open and Efficient Foundation Language Models			2023	

- All leading models use retrieval
- Retrieval models often have to the order O(10^9) parameters
- Much more efficient than largest LLMs

Recap

- **Retrieval-Augmented language models:**
  - Infuse knowledge from external sources into LMs.
  - Suitable for knowledge-intensive tasks where factual accuracy is needed.
- **Main components:** type of external knowledge, type of the LM, type of training.
- Using external knowledge can allow us to reduce the # of parameters of LMs, making them smaller in size without compromising performance.
- **In the LLMs era**:
  - Retrieval aims to augment the prompt.
  - Models are interacting with various tools and APIs to enhance their reasoning capabilities.

References

- Karpukhin, Vladimir, et al. "Dense passage retrieval for open-domain question answering." arXiv preprint arXiv:2004.04906 (2020).
- Guu, Kelvin, et al. "Retrieval augmented language model pre-training." International conference on machine learning. PMLR, 2020.
- Lewis, Patrick, et al. "Retrieval-augmented generation for knowledge-intensive nlp tasks." Advances in Neural Information Processing Systems 33 (2020): 9459-9474.
- Schick, Timo, et al. "Toolformer: Language models can teach themselves to use tools." arXiv preprint arXiv:2302.04761 (2023).
- Yao, Shunyu, et al. "React: Synergizing reasoning and acting in language models." arXiv preprint arXiv:2210.03629 (2022).
- Mialon, Grégoire, et al. "Augmented language models: a survey." arXiv preprint arXiv:2302.07842 (2023).