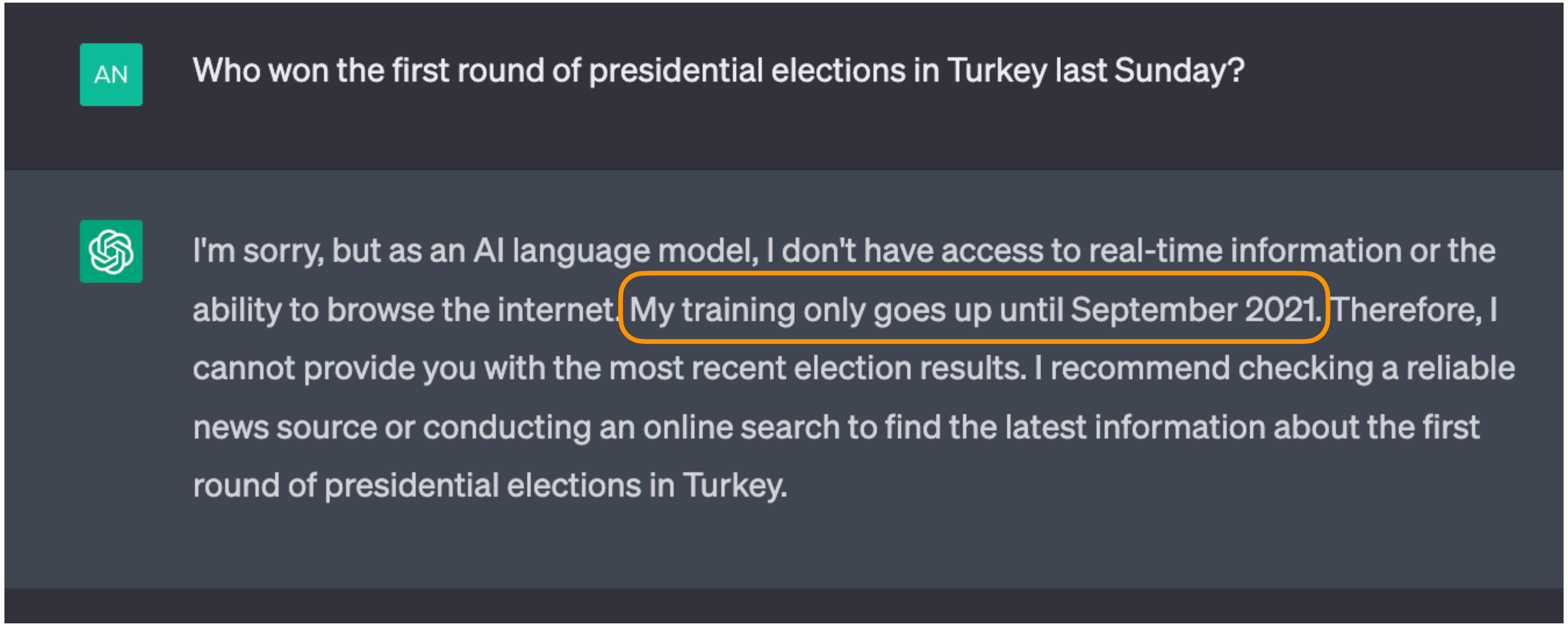
Retrieval-Augmented LMs

Angelika Romanou

Reading Comprehension Challenges



Who won the first round of presidential elections in Turkey last Sunday?

I'm sorry, but as an AI language model, I don't have access to real-time information or the ability to browse the internet. My training only goes up until September 2021. Therefore, I cannot provide you with the most recent election results. I recommend checking a reliable news source or conducting an online search to find the latest information about the first round of presidential elections in Turkey.

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

Limitations of PLMs (& LLMs)

- Hallucination problem
- Struggle to apply precise knowledge
- Cannot easily expand or update their parameters on inference time

What tools give us direct access to information?

Why retrieval is good

- Precise knowledge access mechanism
- Easy update on test time
- Neural Retrieval

Limitation

Task-specific way to integrate into downstream tasks

Today's Outline

- **Lecture: Retrieval-Augmented LMs**
  - **Aspects of Retrieval-Augmented LMs:** Model types, training objectives, different external knowledge
  - **Downstream tasks:** Tasks & Dataset
  - **Augmented LLMs:** Retrieval in the LLM era
  - **Augmentation benefits:** Explainability, Modularity, Parameter efficiency

Why we cannot do that with
Extractive or Generative QA models?

Finding the answer in 21M documents

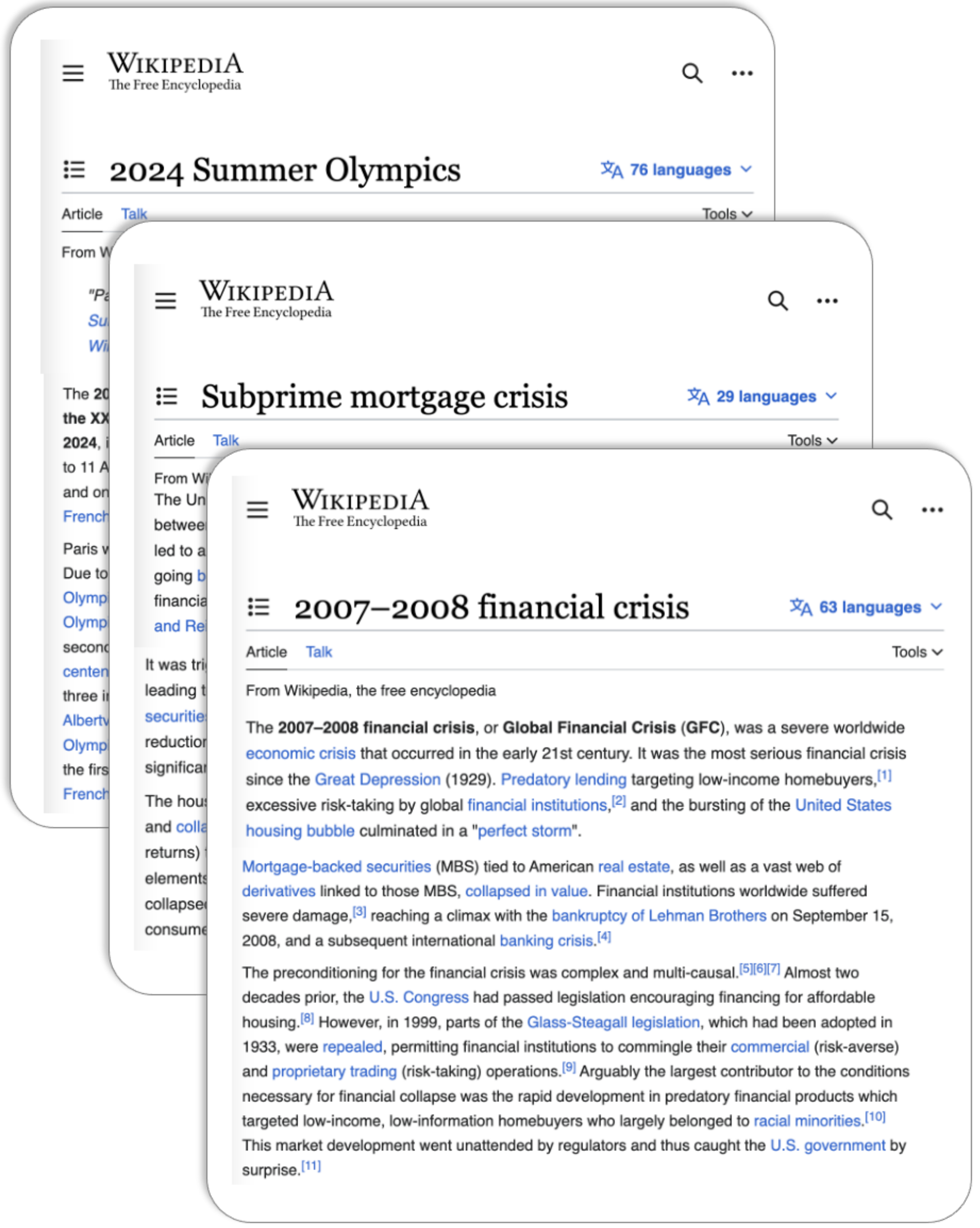
Query

Documents

Retrieve relevant documents

That might contain the answer

"Where the financial crisis of 2008 started?"



WIKIPEDIA
The Free Encyclopedia

2024 Summer Olympics — 76 languages

Article Talk Tools

WIKIPEDIA
The Free Encyclopedia

Subprime mortgage crisis — 29 languages

Article Talk Tools

WIKIPEDIA
The Free Encyclopedia

2007–2008 financial crisis — 63 languages

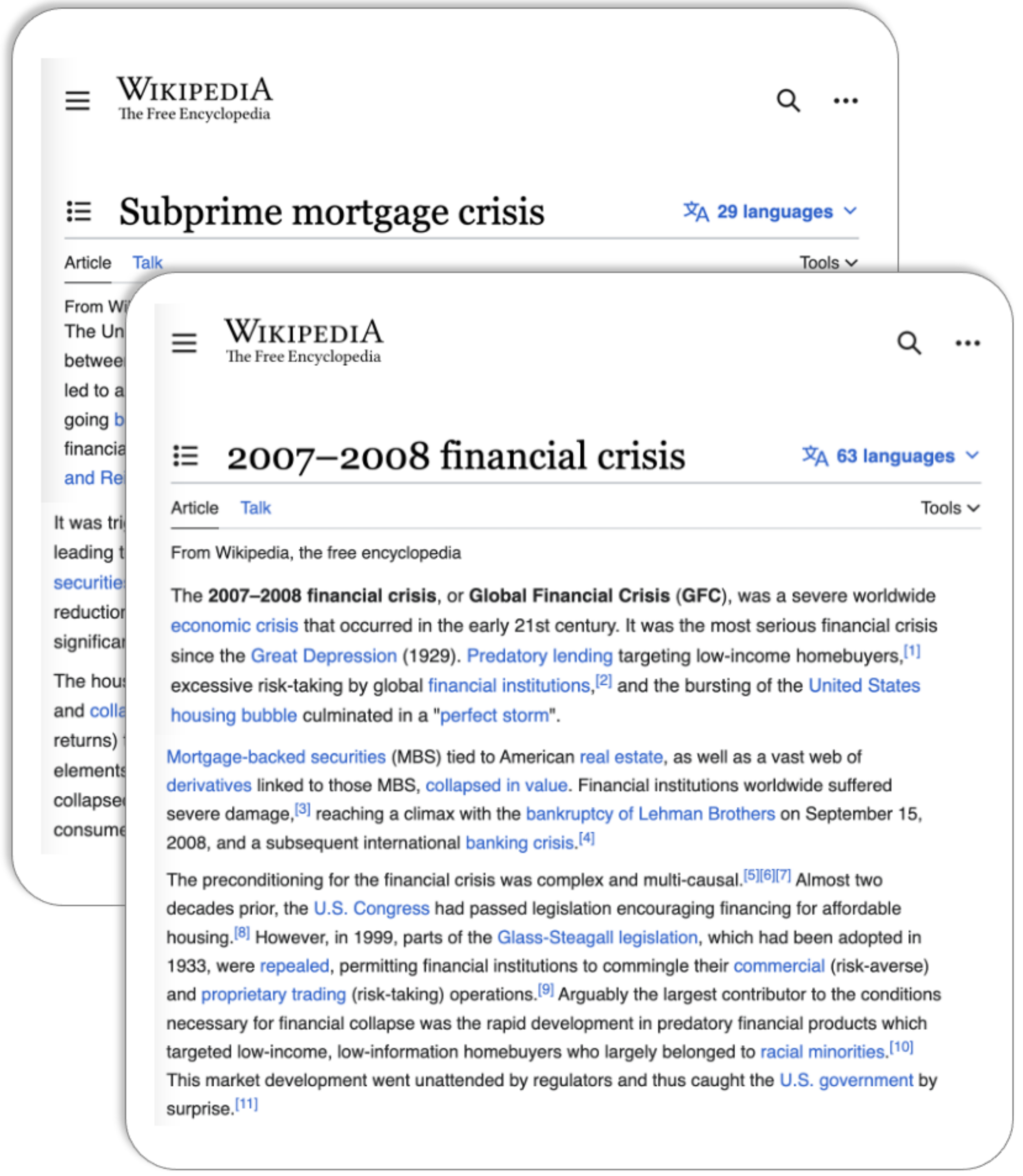
Article Talk Tools

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]



WIKIPEDIA
The Free Encyclopedia

Subprime mortgage crisis — 29 languages

Article Talk Tools

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]

Dense Passage Retrieval (DPR)

Documents

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]

Dense Embedding Model

Query

"Where the financial crisis of 2008 started?"

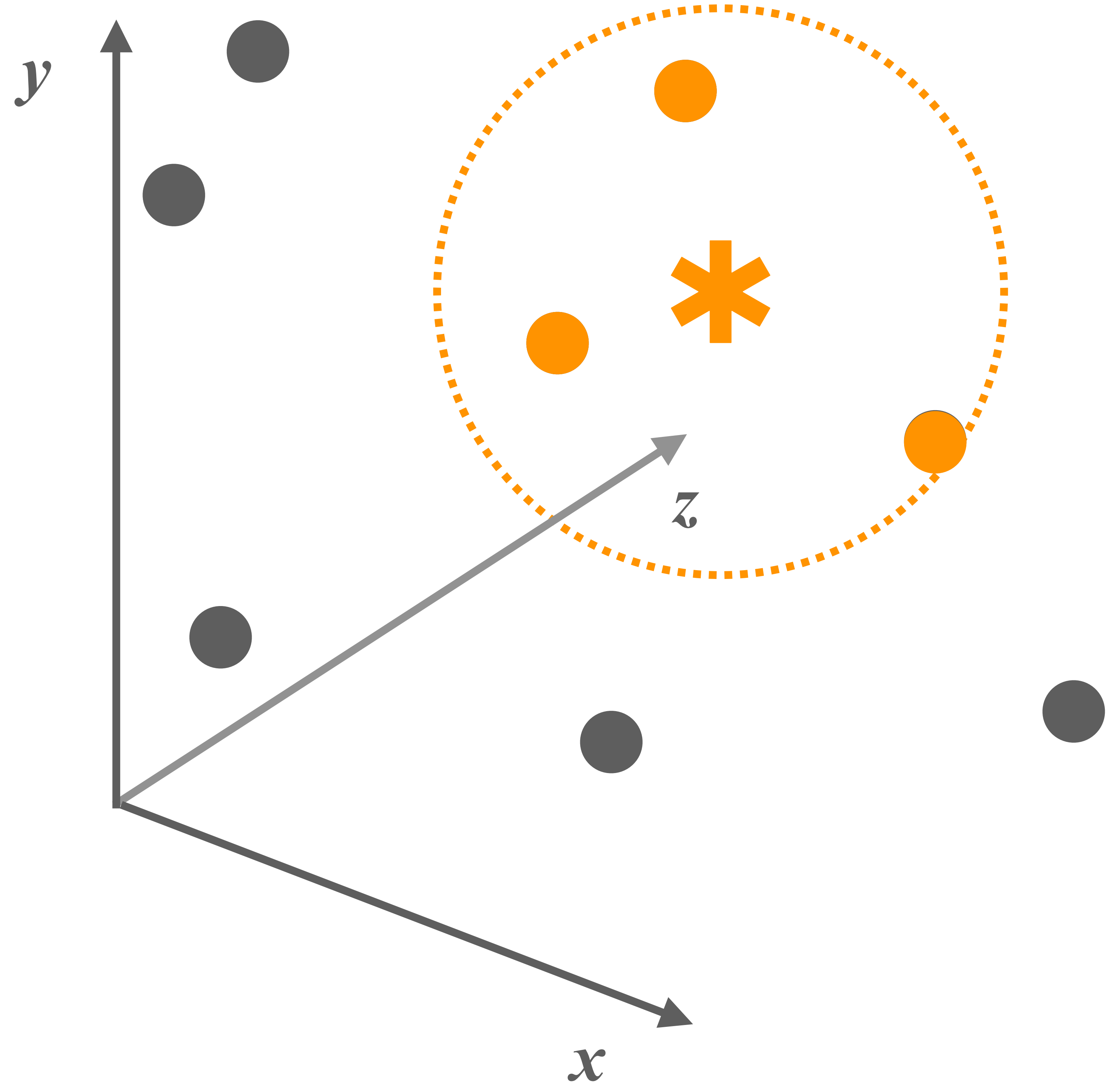
- Create the representations of documents
- Create the representation of the query
- Retrieve *k* documents vectors based on the query vector

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, ...]

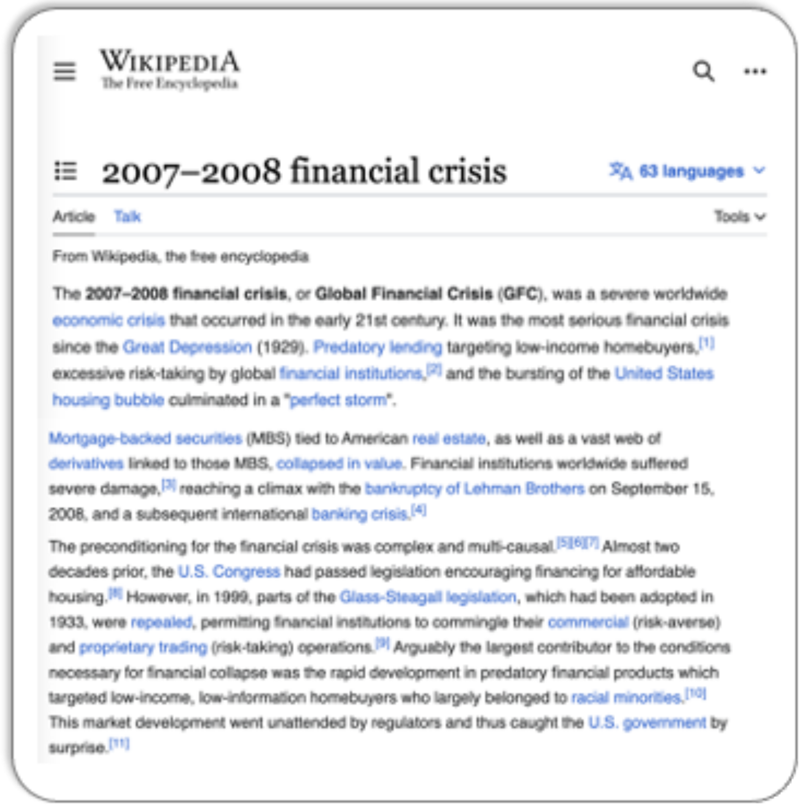
Training DPR

How to create a Document-Query vector space?

Goal: **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

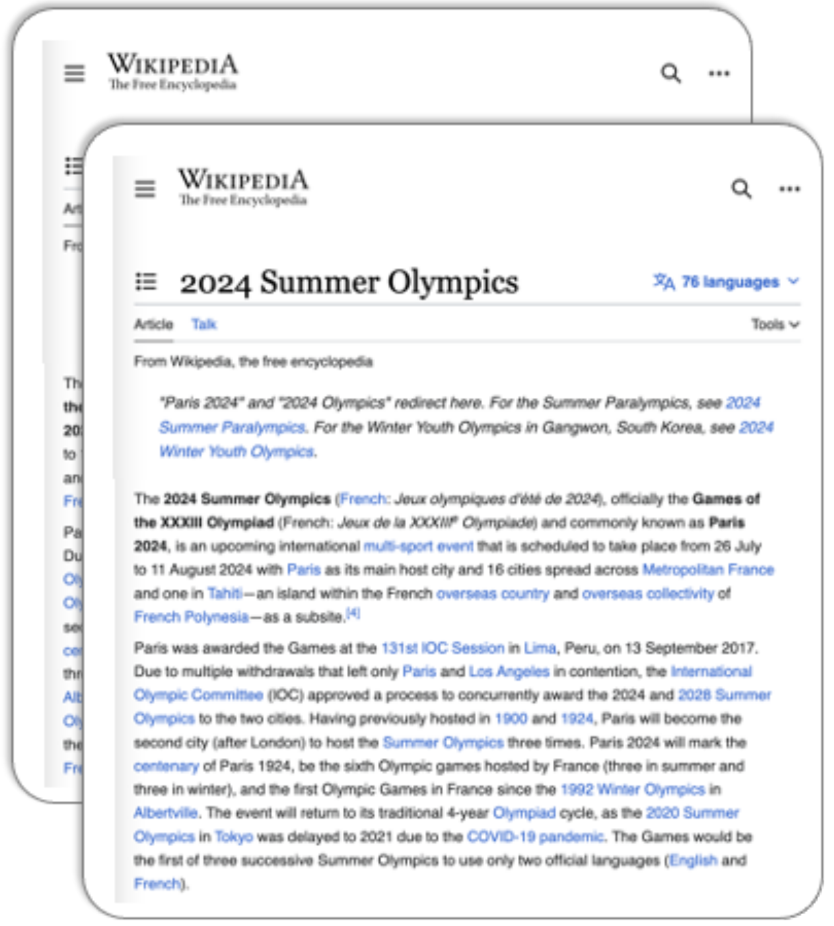
From Wikipedia, the free encyclopedia

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 passages 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 XXXIII^e 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 can we integrate a neural retriever into a Language Model?

Retrieval-Augmented LMs

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

LM

Retriever

Auto-Encoder

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

$$p(z \mid x)$$

Auto-Encoder

Auto-Regressive

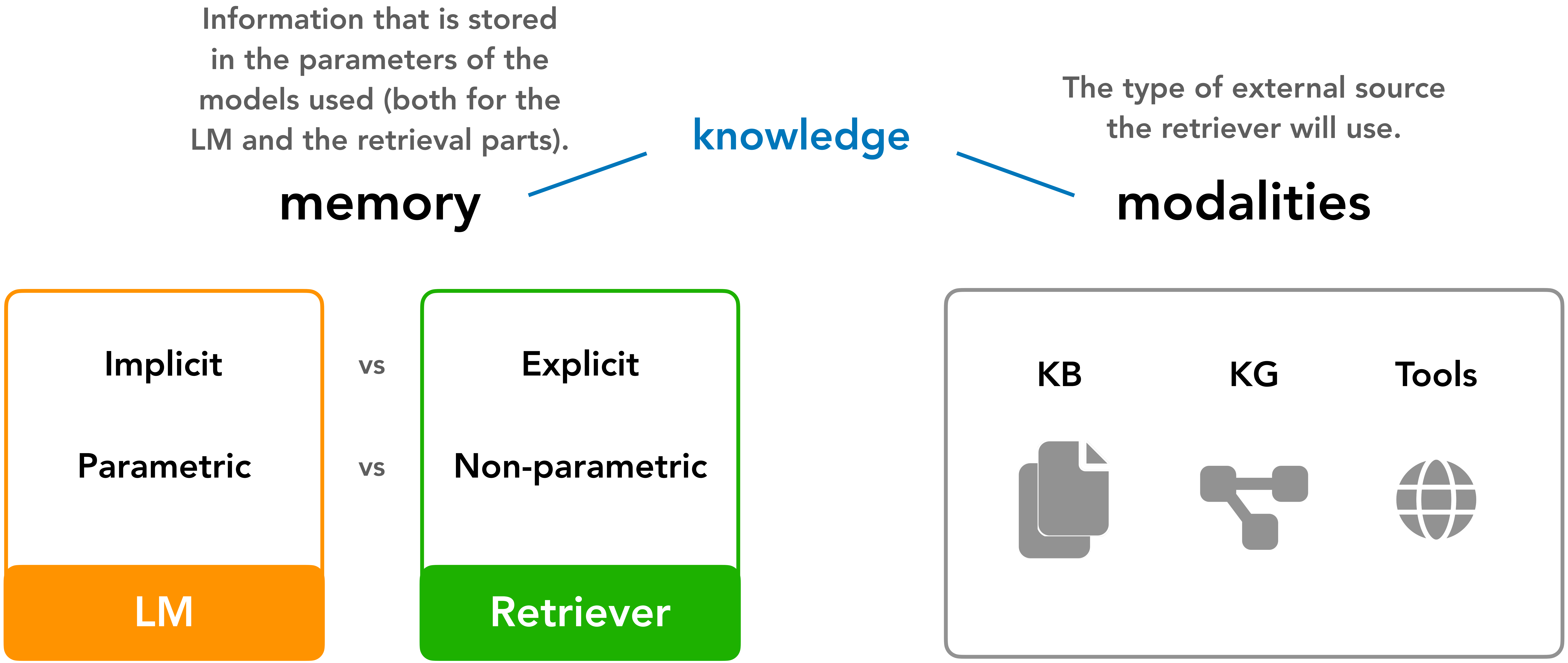
$$\sum_{z \in \mathcal{Z}} \prod_{i}^{N} \quad p(y_i \mid x, z, y_{1:i-1})$$

Trained over what documents are relevant and should be retrieved.

z is a latent variable that corresponds to the retrieved documents.

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

The type of external source the retriever will use.

modalities

Implicit vs Explicit

Parametric vs Non-parametric

LM

Retriever

KB

KG

Tools

The landscape of Retrieval-Augmented LMs

ARCHITECTURE OF THE LM

Generative

vs

Extractive

RAG: Fine-tuning & KB

TRAINING OF THE COMPONENTS

Pre-training

vs

Fine-tuning

REALM: Pre-training & KB

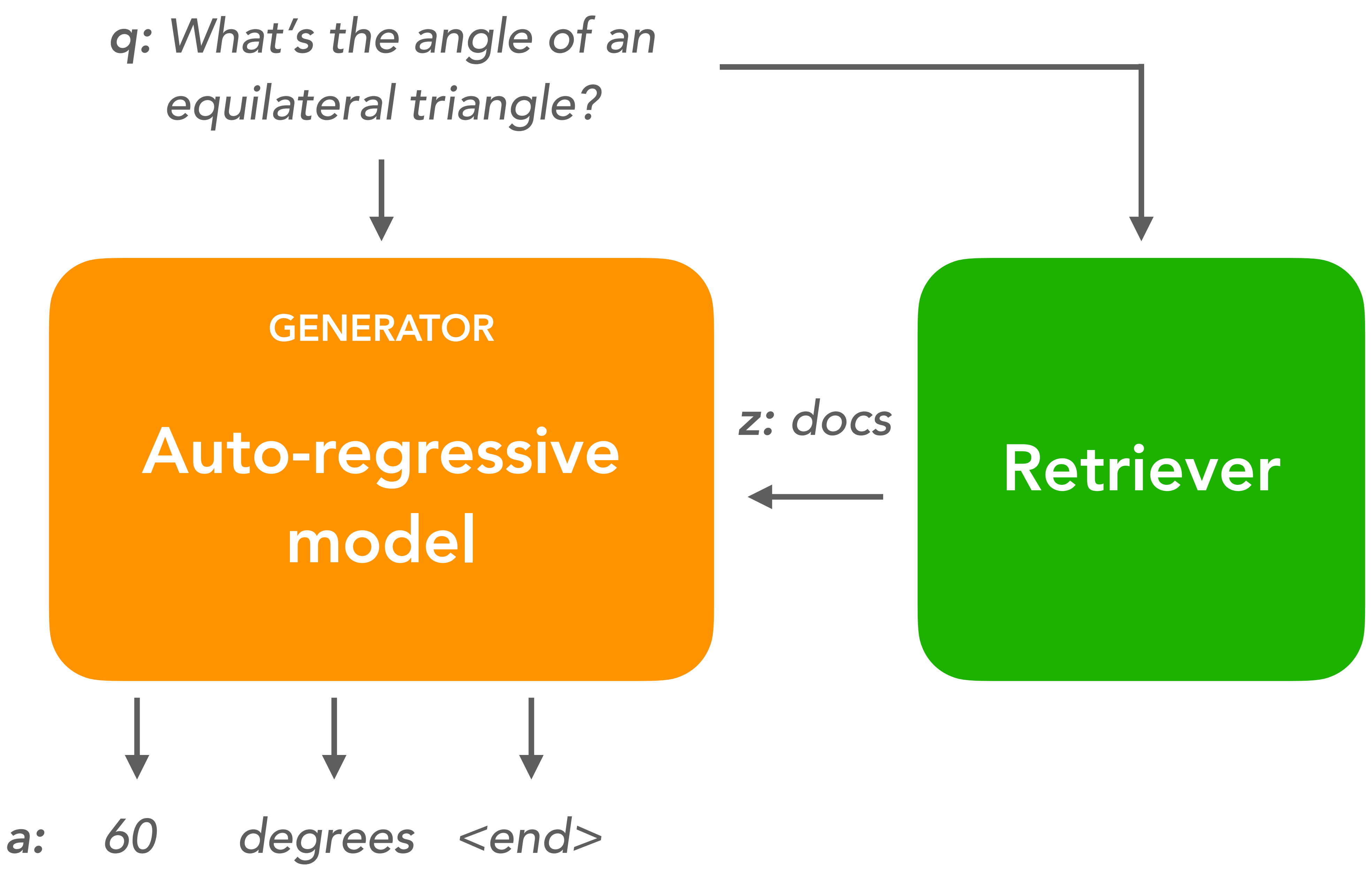
TYPES OF EXTERNAL KNOWLEDGE

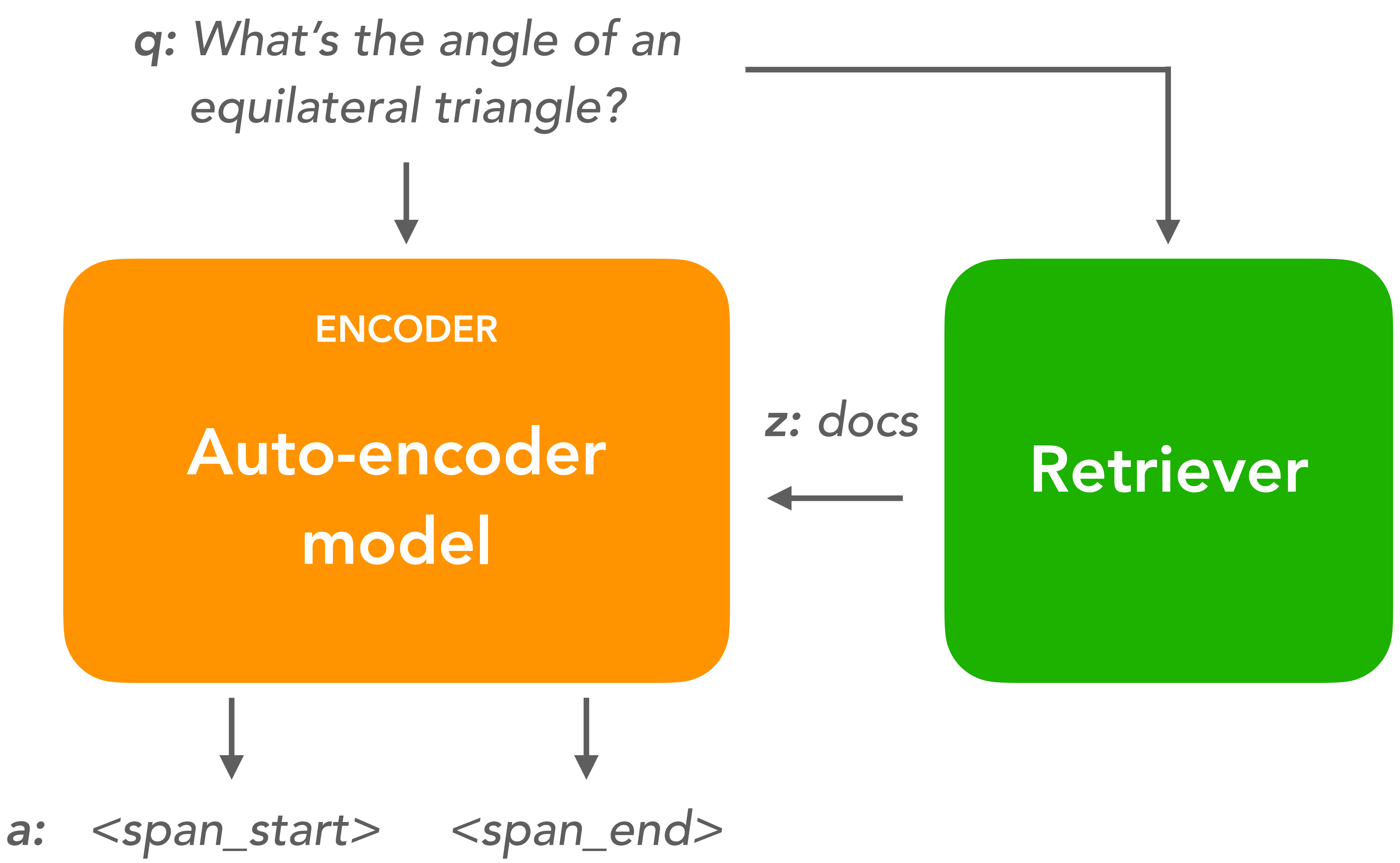
Knowledge Bases

Knowledge Graphs

ERNIE: Pre-training & KG

Generative vs Extractive

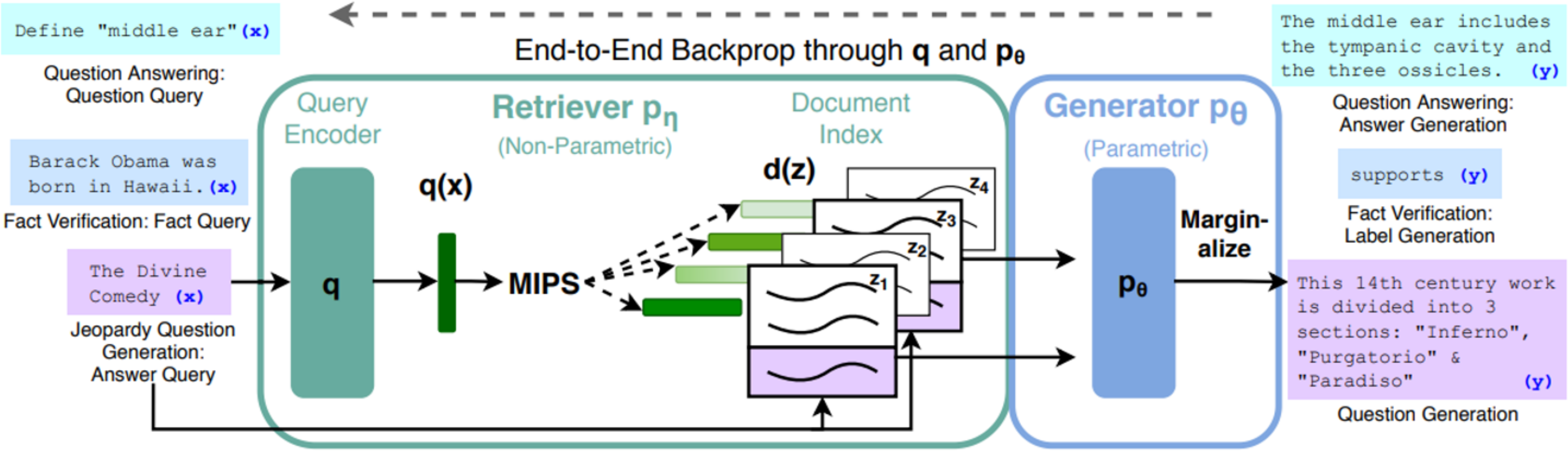




RAG: Generative Retrieval-Augmented LM

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

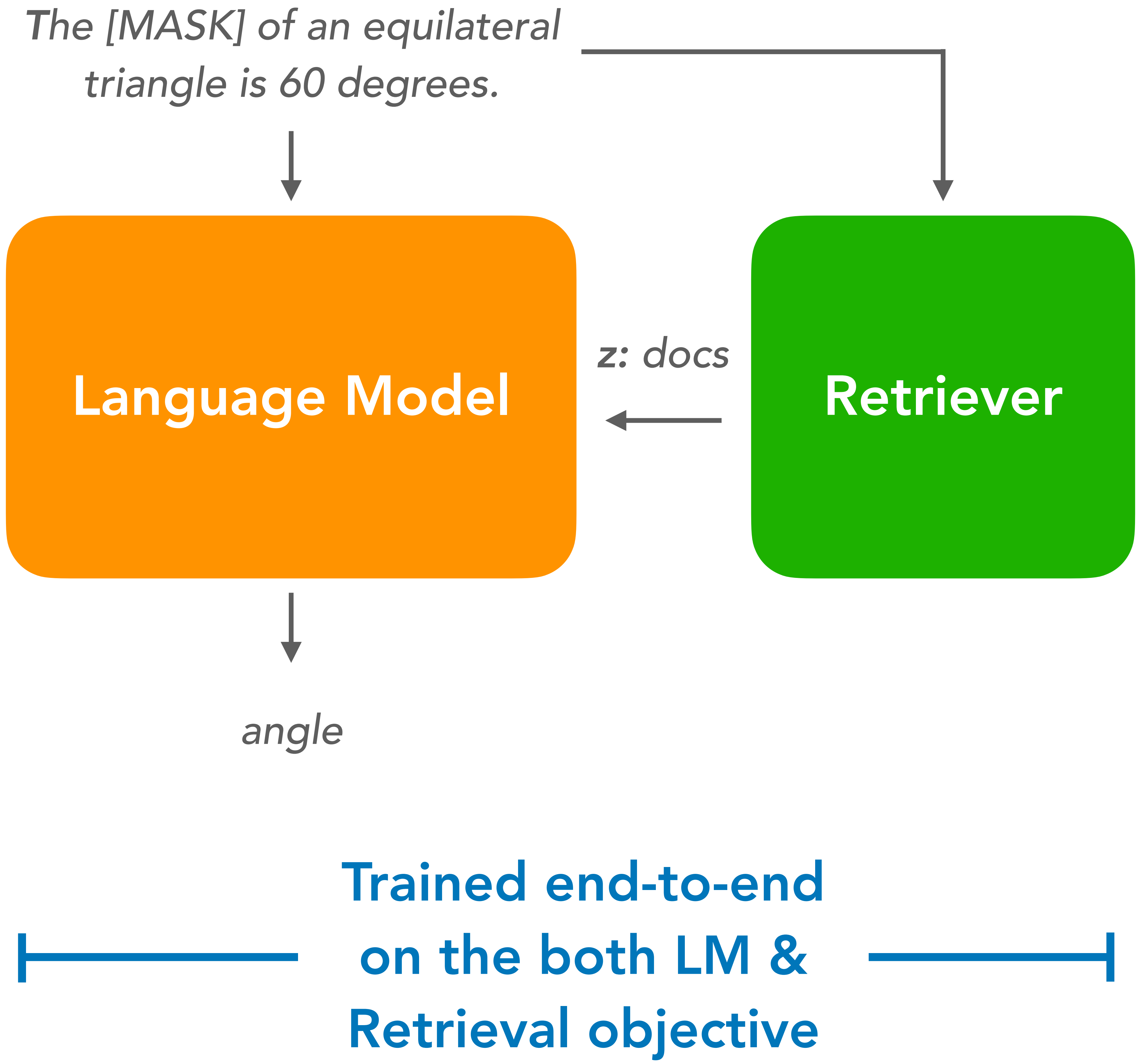
$$p(y \mid x) = \sum_{z \in \mathcal{Z}} \prod_{i}^{N} p(y_i|x, z, y_{1:i-1})$$



Define "middle ear" (x)
Question Answering: Question Query
Barack Obama was born in Hawaii. (x)
Fact Verification: Fact Query
The Divine Comedy (x)
Jeopardy Question Generation: Answer Query
End-to-End Backprop through q and pθ
Query Encoder
q
q(x)
Retriever pη
(Non-Parametric)
MIPS
Document Index
d(z)
z4
z3
z2
z1
Generator pθ
(Parametric)
pθ
Margin-alize
The middle ear includes the tympanic cavity and the three ossicles. (y)
Question Answering: Answer Generation
supports (y)
Fact Verification: Label Generation
This 14th century work is divided into 3 sections: "Inferno", "Purgatorio" & "Paradiso" (y)
Question Generation

Lewis et al. (2020)

Pre-training vs Fine-tuning



The [MASK] of an equilateral triangle is 60 degrees.

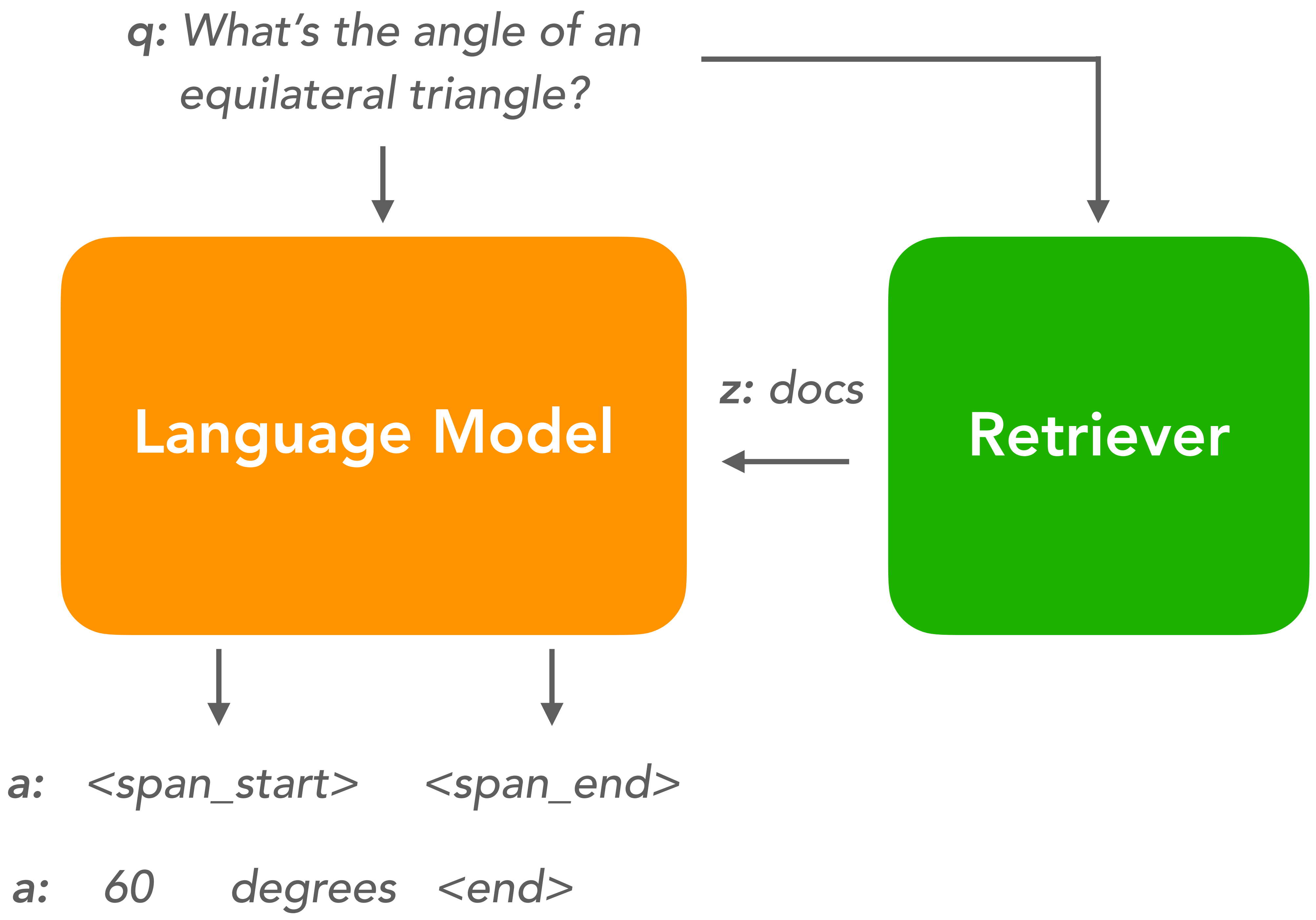
Language Model

z: *docs*

Retriever

angle

Trained end-to-end on the both LM & Retrieval objective



q: *What's the angle of an equilateral triangle?*

Language Model

z: *docs*

Retriever

a: *<span_start>* *<span_end>*

a: *60* *degrees* *<end>*

REALM: Pre-training Retrieval Augmented LMs

First Retrieve:

The retriever model is trained on what documents are relevant.

Goal: Penalise uninformative retrievals

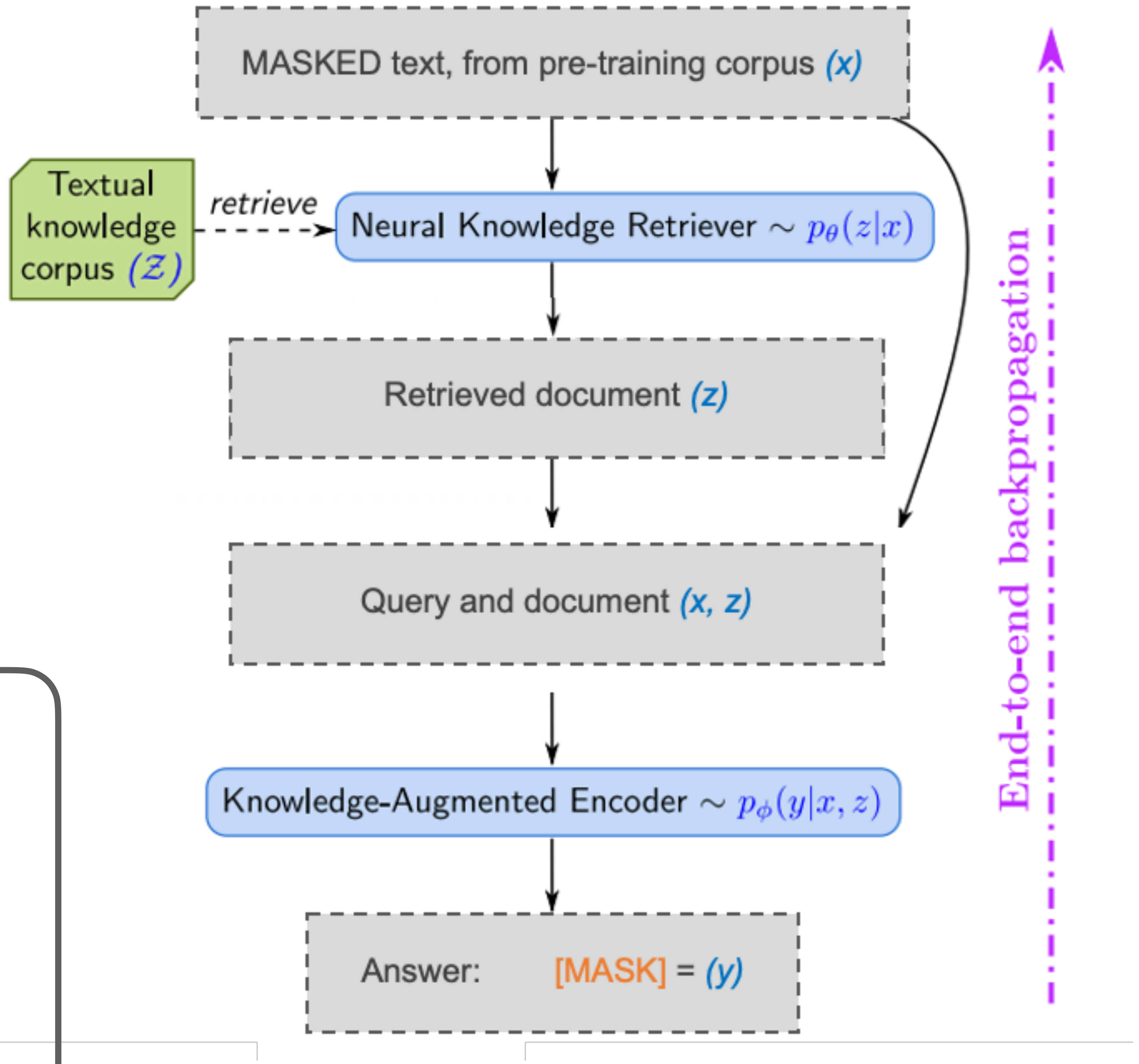
Then Predict:

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

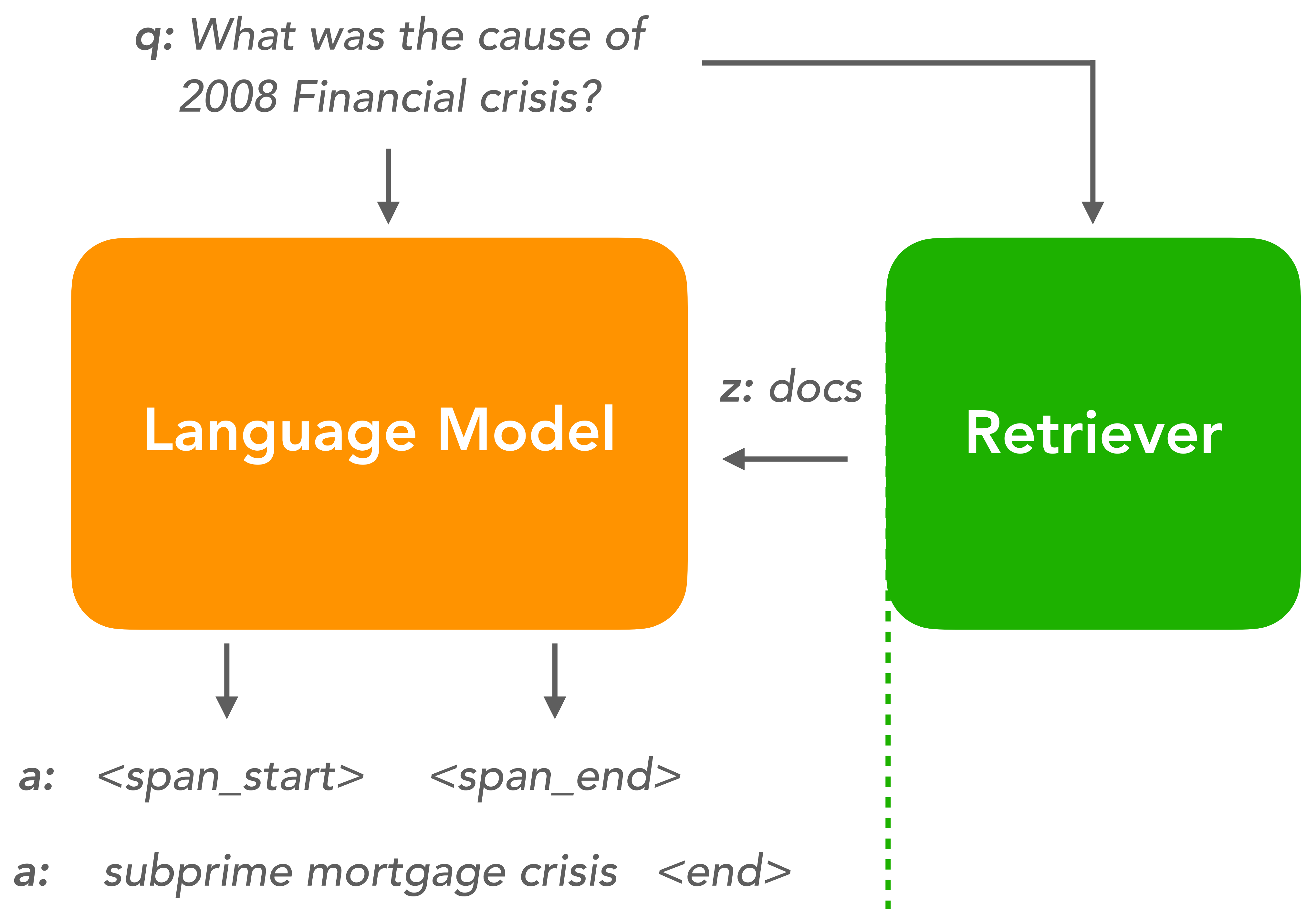
Benefits of pre-training end-to-end

- The retriever is trained to fetch documents that minimize perplexity.
- Model-centric **unsupervised alignments** between text in the pre-training corpus and knowledge corpus.


MASKED text, from pre-training corpus (x)
Textual knowledge corpus ($\mathcal{Z}$)
retrieve
Neural Knowledge Retriever $\sim p_\theta(z|x)$
Retrieved document (z)
Query and document (x, z)
Knowledge-Augmented Encoder $\sim p_\phi(y|x,z)$
Answer: [MASK] = (y)
End-to-end backpropagation

Guu et al. (2020)

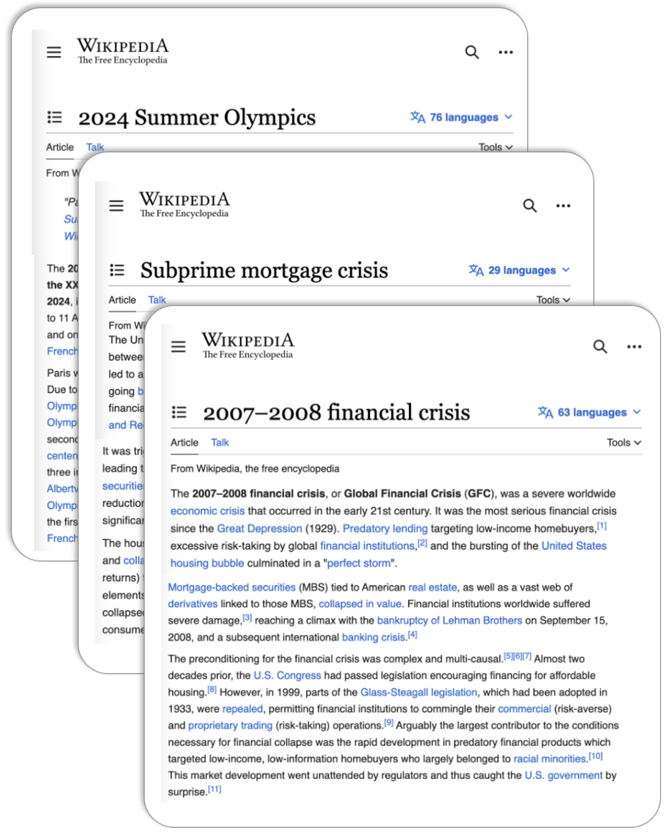
Different types of external knowledge


q: What was the cause of 2008 Financial crisis?
Language Model
z: docs
Retriever
a: <span_start> <span_end>
a: subprime mortgage crisis <end>

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

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

                    ...
```



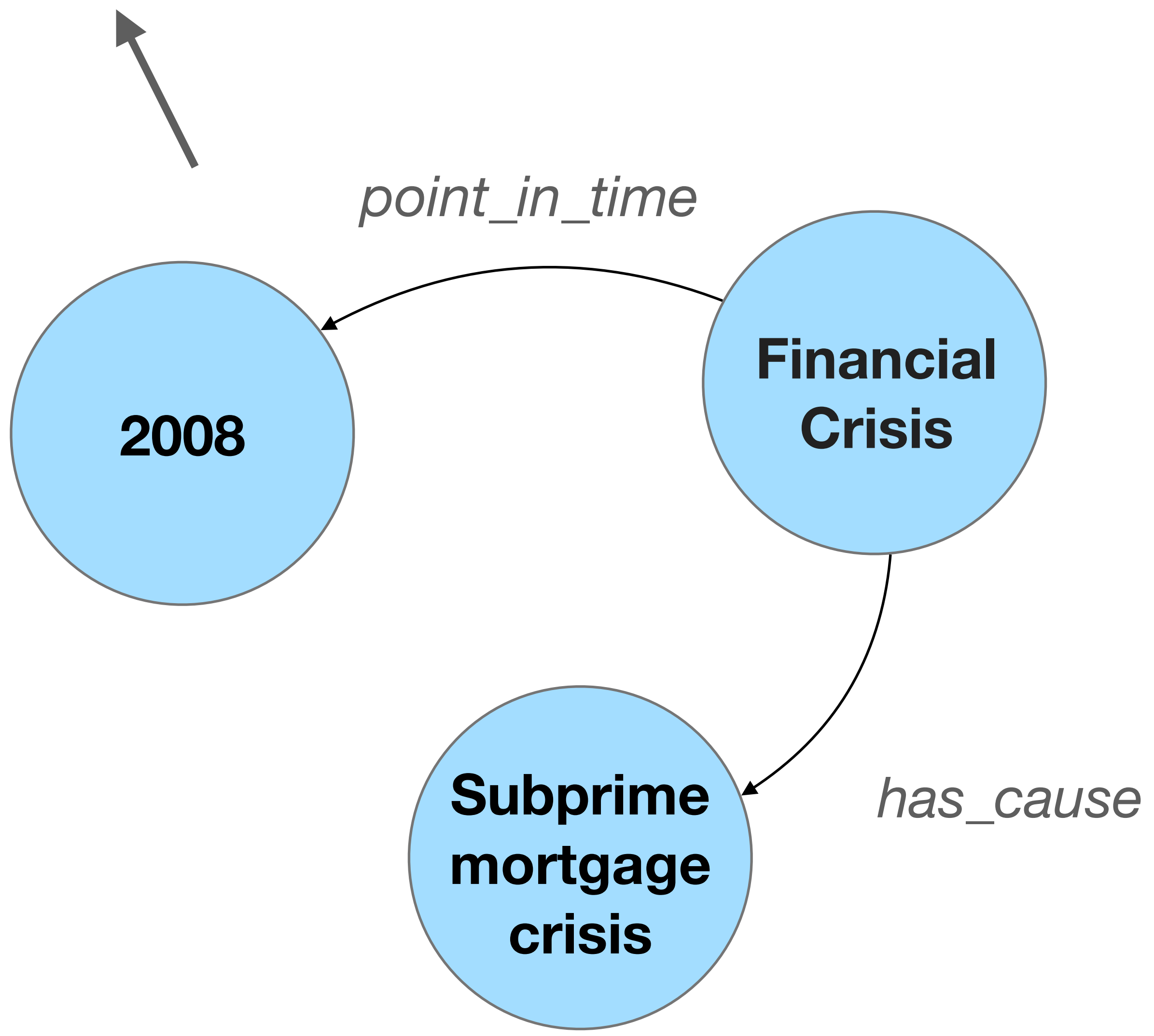
From Wikipedia, the free encyclopedia

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]

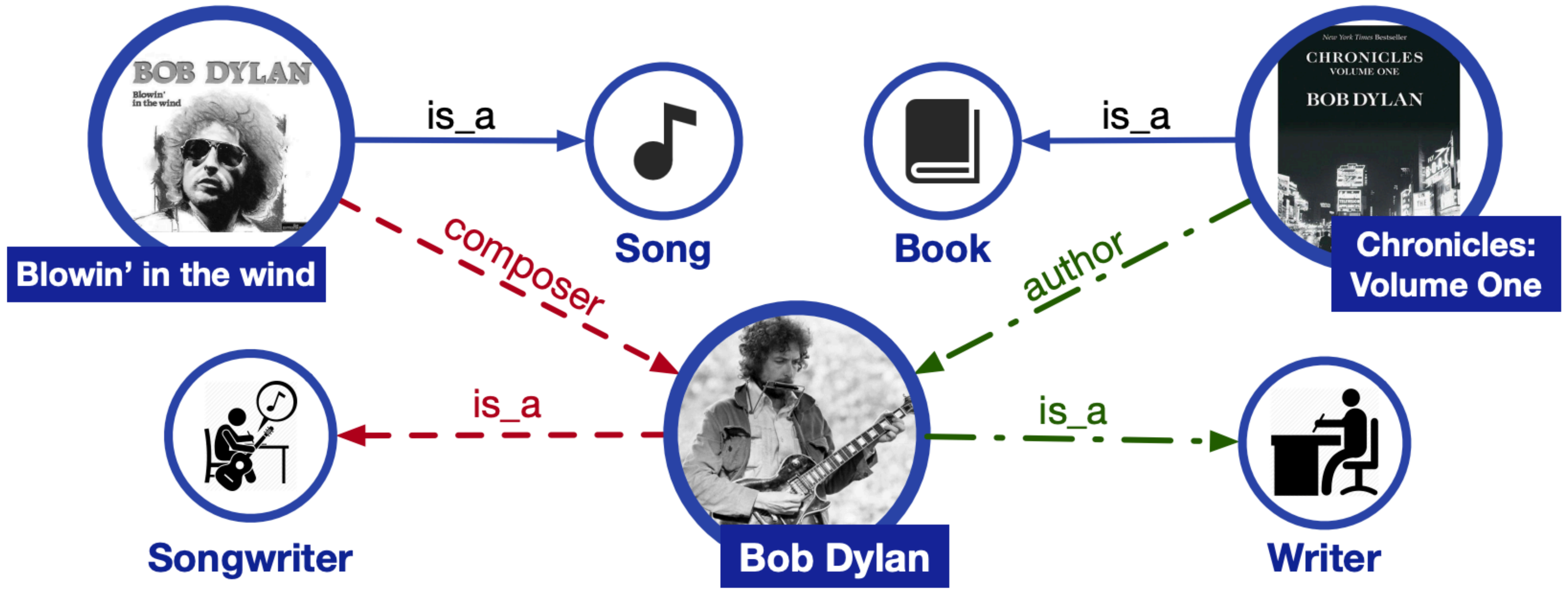
```
[Financial Crisis, point_in_time, 2008]
[Financial Crisis, has cause, subprime mortgage crisis]
```



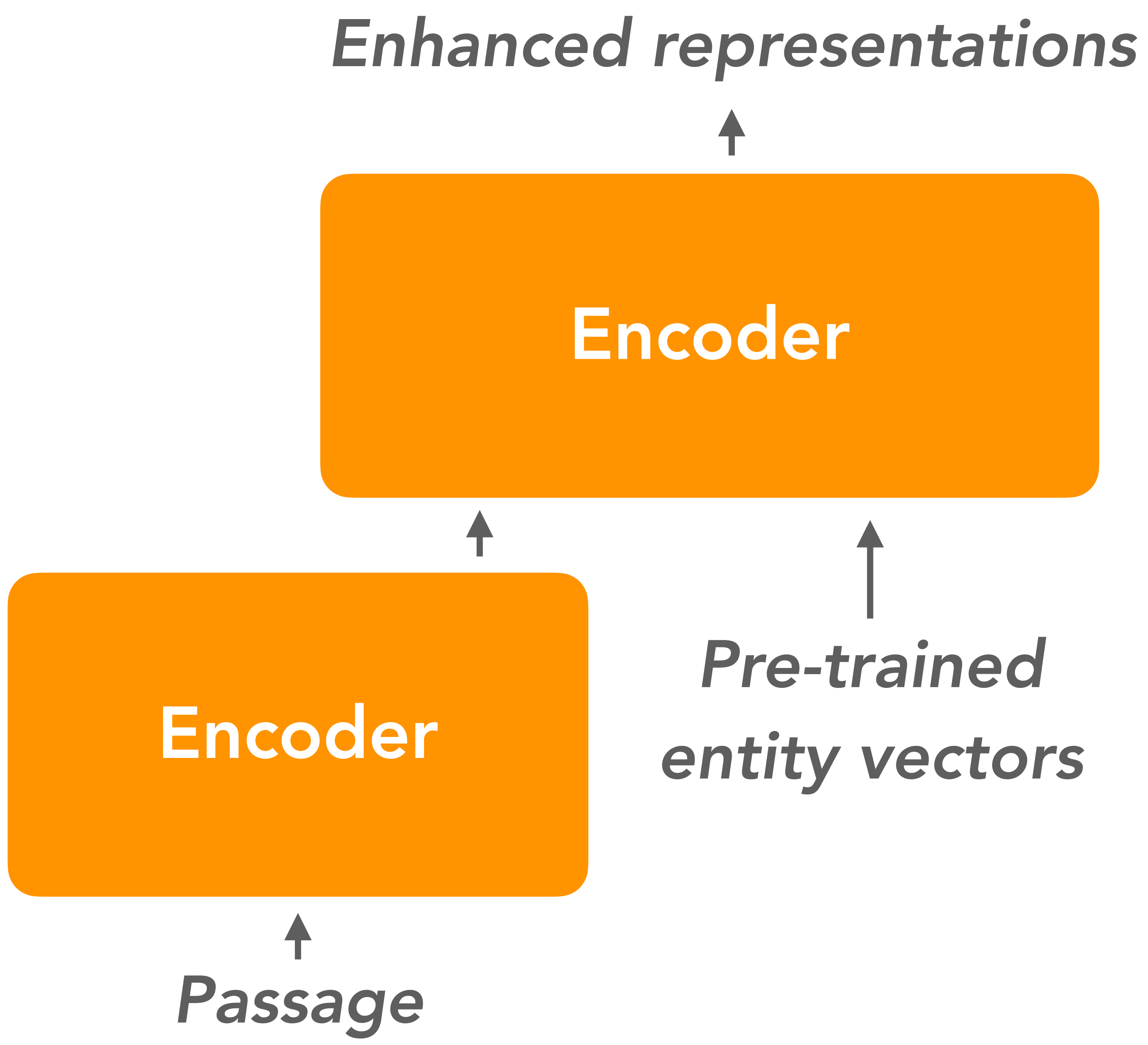
The retriever aims to create a shared vector space between the used modality & the text in the input query.

ERNIE: Infuse KG knowledge

- Knowledge graphs contain rich **structured** knowledge facts
- Integrating KG information can be challenging:
  - How do we extract and encode the information in the KG?
  - How do we do information fusion between these heterogenous modalities?



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



Zhang et al. (2019)

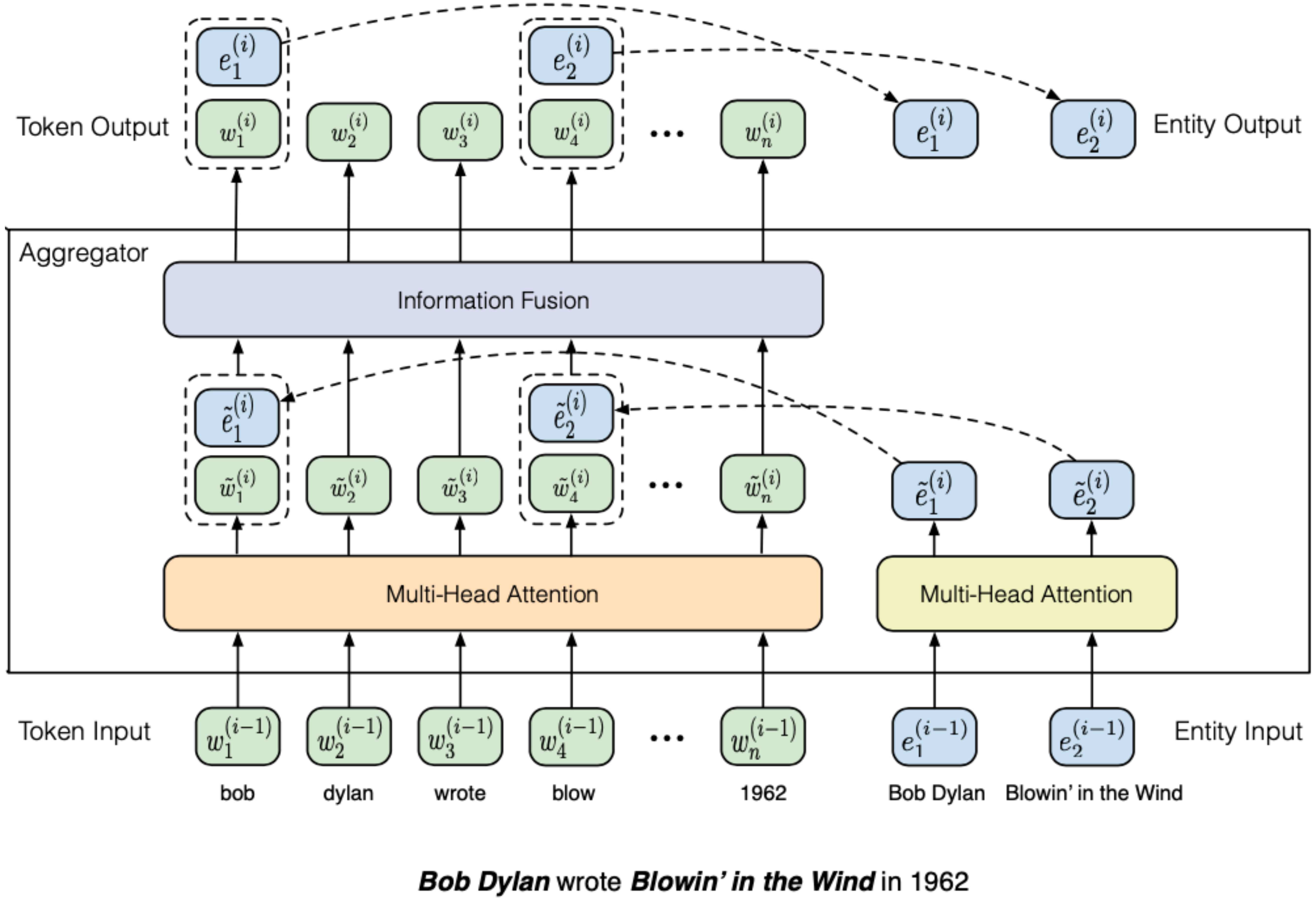
ERNIE: Infuse KG knowledge

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.



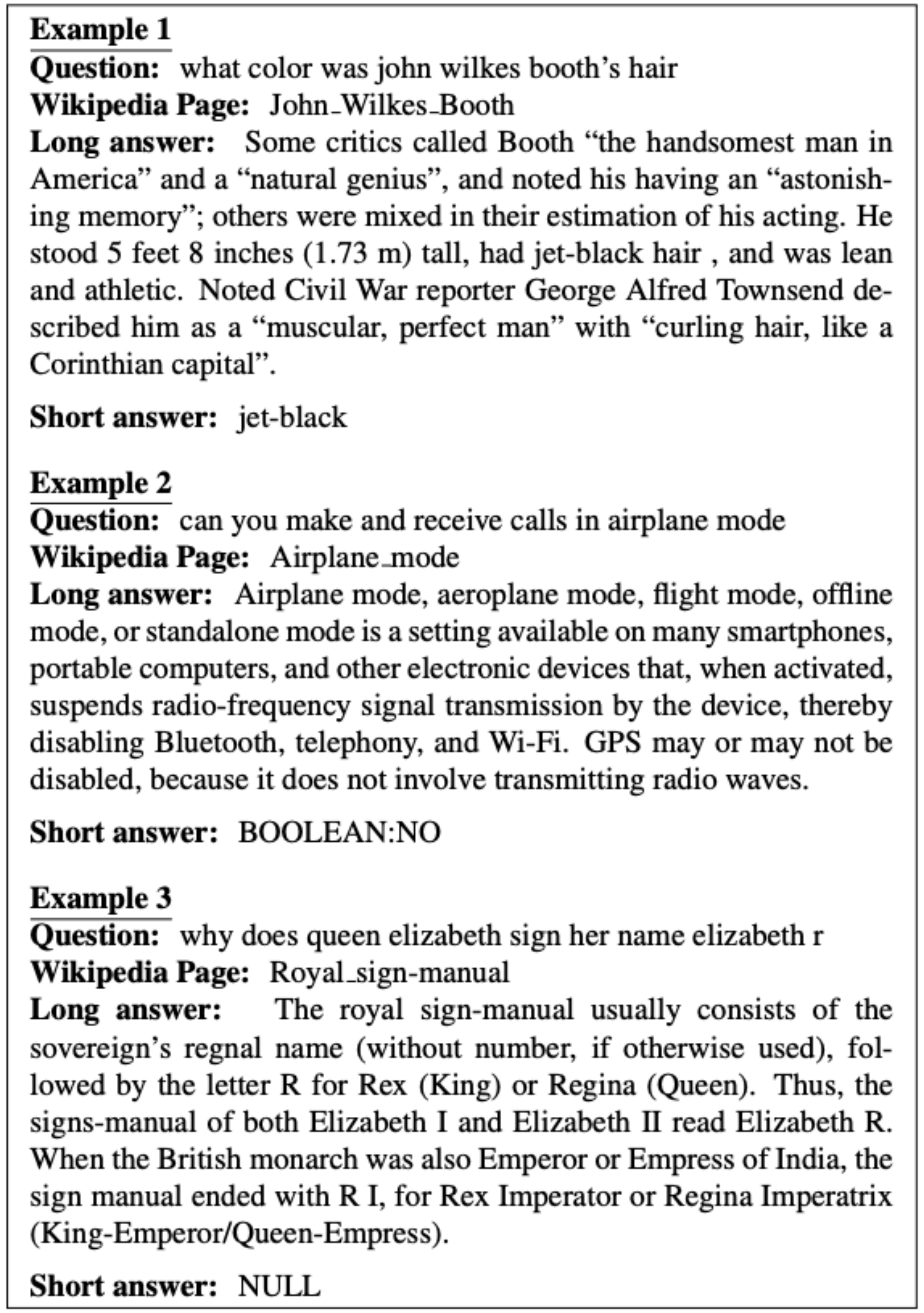
Zhang et al. (2019)

Factual-heavy NLP Datasets: Natural Questions

- Contains both long-form & short-form answers.
- The **questions** consist of real anonymized, aggregated queries issued to the **Google search** engine.
- The questions (Google queries) are not similar to the text containing the answer (Wikipedia).
- The span that contains the answer can be located **across sentences**.

Train split 300K

Test split 15K



Example 1
Question: what color was john wilkes booth's hair
Wikipedia Page: John_Wilkes_Booth
Long answer: Some critics called Booth "the handsomest man in America" and a "natural genius", and noted his having an "astonishing memory"; others were mixed in their estimation of his acting. He stood 5 feet 8 inches (1.73 m) tall, had jet-black hair , and was lean and athletic. Noted Civil War reporter George Alfred Townsend described him as a "muscular, perfect man" with "curling hair, like a Corinthian capital".

Short answer: jet-black

Example 2
Question: can you make and receive calls in airplane mode
Wikipedia Page: Airplane_mode
Long answer: Airplane mode, aeroplane mode, flight mode, offline mode, or standalone mode is a setting available on many smartphones, portable computers, and other electronic devices that, when activated, suspends radio-frequency signal transmission by the device, thereby disabling Bluetooth, telephony, and Wi-Fi. GPS may or may not be disabled, because it does not involve transmitting radio waves.

Short answer: BOOLEAN:NO

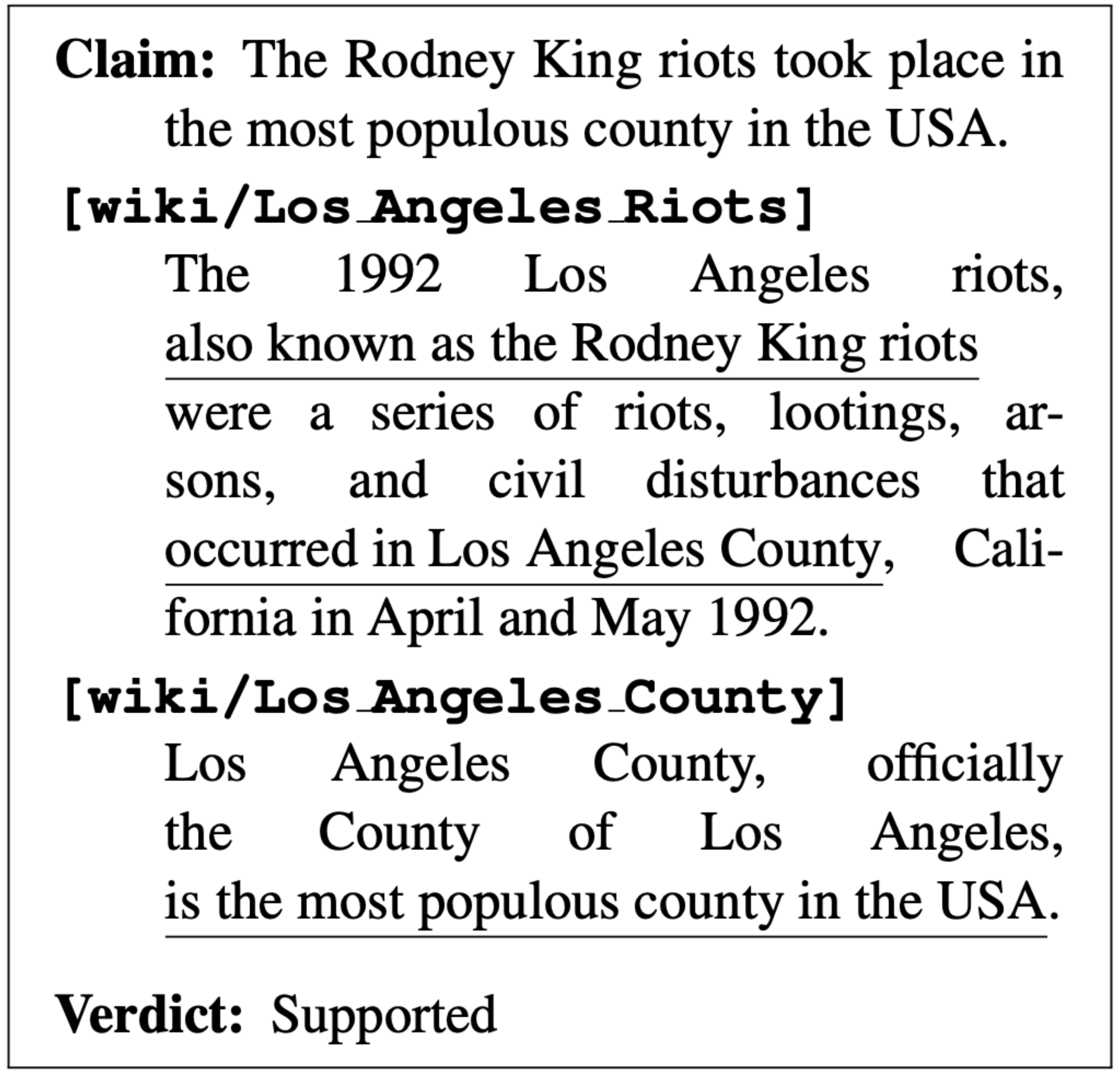
Example 3
Question: why does queen elizabeth sign her name elizabeth r
Wikipedia Page: Royal_sign-manual
Long answer: The royal sign-manual usually consists of the sovereign's regnal name (without number, if otherwise used), followed by the letter R for Rex (King) or Regina (Queen). Thus, the signs-manual of both Elizabeth I and Elizabeth II read Elizabeth R. When the British monarch was also Emperor or Empress of India, the sign manual ended with R I, for Rex Imperator or Regina Imperatrix (King-Emperor/Queen-Empress).

Short answer: NULL

Kwiatkowski et al. (2019)

Factual-heavy NLP datasets: FEVER

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.

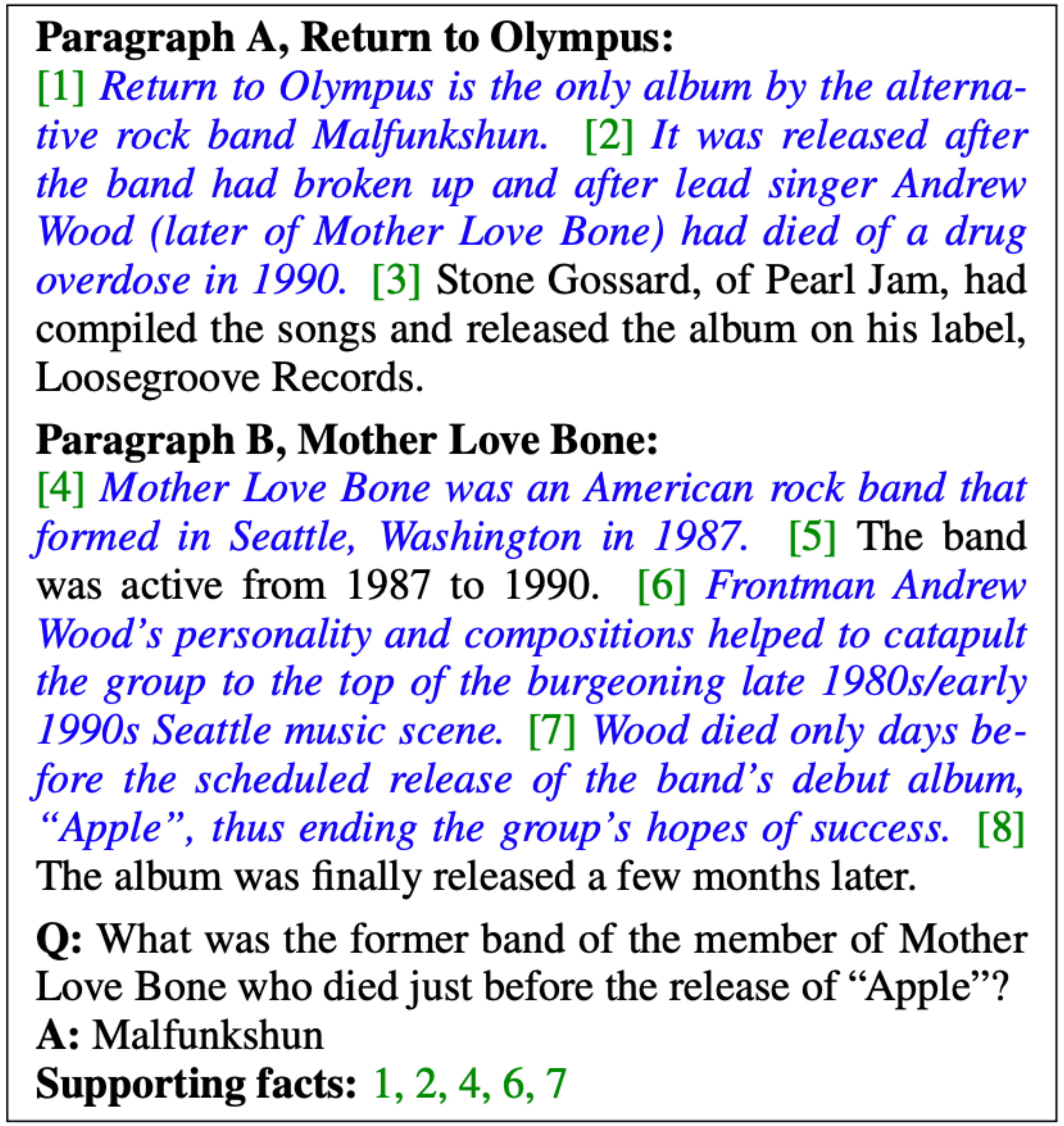
Verdict: Supported

- It consists of 185K claims generated by altering sentences extracted from Wikipedia.
- The claims are classified as SUPPORTED, REFUTED, or NOTENOUGHINFO.
- For the first two classes, the dataset provides the **pieces of evidence supporting or refuting the claim**.

Thorne et al. (2018)

Factual-heavy NLP datasets: HotpotQA

Multi-hop QA



Paragraph A, Return to Olympus:
[1] *Return to Olympus is the only album by the alternative rock band Malfunkshun.* [2] *It was released after the band had broken up and after lead singer Andrew Wood (later of Mother Love Bone) had died of a drug overdose in 1990.* [3] Stone Gossard, of Pearl Jam, had compiled the songs and released the album on his label, Loosegroove Records.

Paragraph B, Mother Love Bone:
[4] *Mother Love Bone was an American rock band that formed in Seattle, Washington in 1987.* [5] The band was active from 1987 to 1990. [6] *Frontman Andrew Wood's personality and compositions helped to catapult the group to the top of the burgeoning late 1980s/early 1990s Seattle music scene.* [7] *Wood died only days before the scheduled release of the band's debut album, "Apple", thus ending the group's hopes of success.* [8] The album was finally released a few months later.

Q: What was the former band of the member of Mother Love Bone who died just before the release of "Apple"?
A: Malfunkshun
Supporting facts: 1, 2, 4, 6, 7

- The dataset contains questions, answers, and supported facts that the answer is based on.
- The questions require finding and **reasoning over multiple supporting documents to answer**.

Need for complex reasoning to answer the questions.

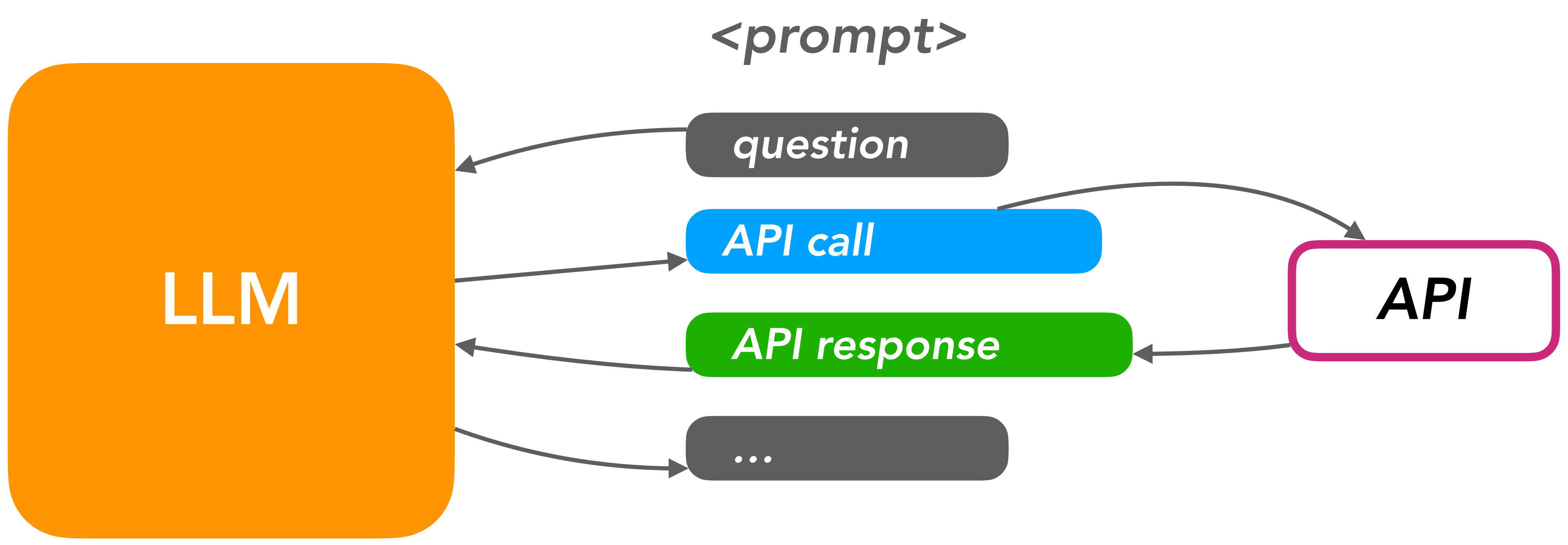
Yang et al. (2018)

How can we retrieval-augment LLMs?

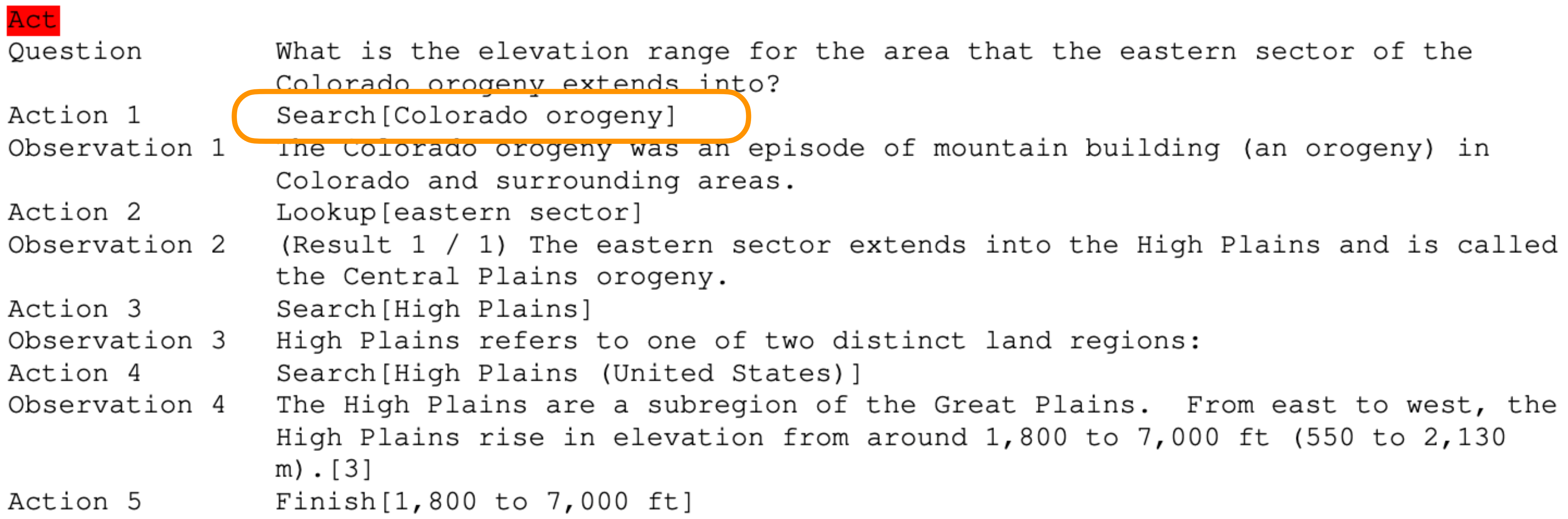
Augmented LLMs

Retrieve from tools & APIs

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

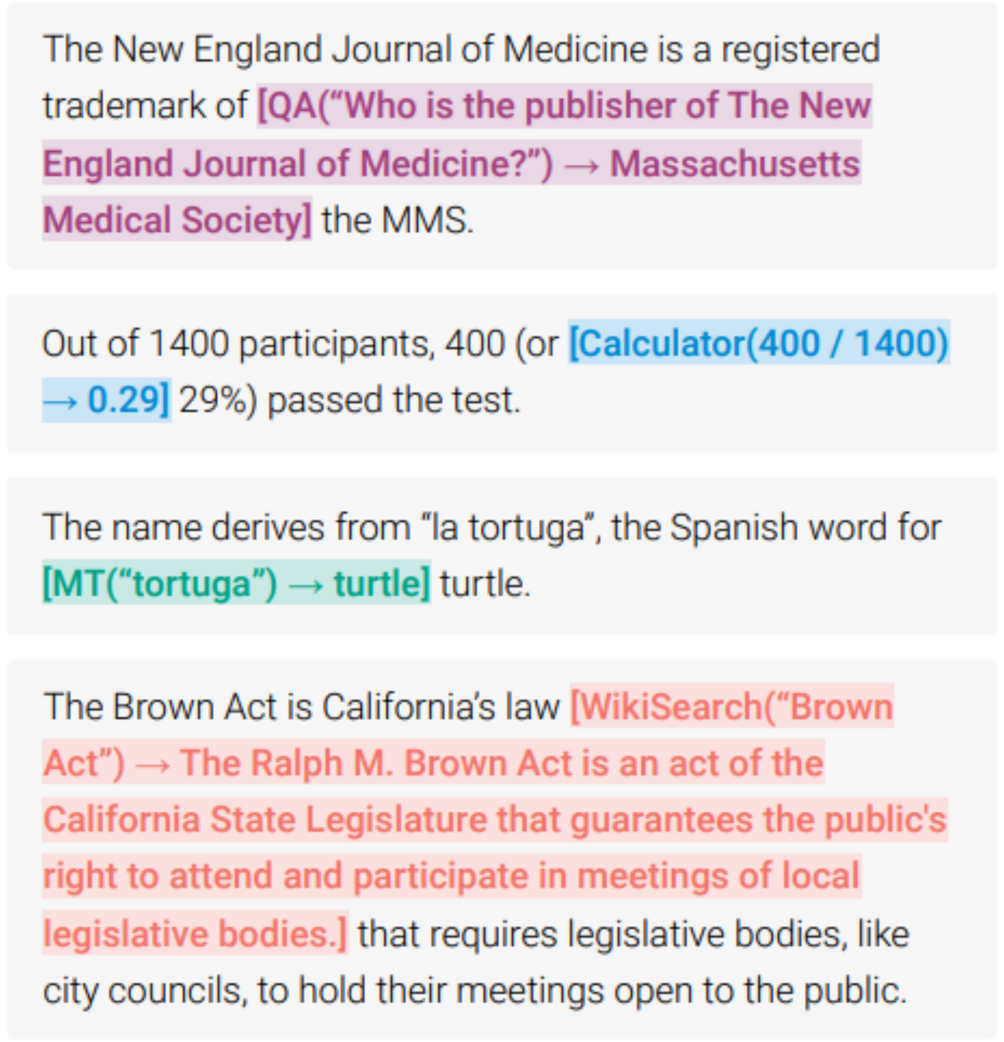


Retrieval-Augmented Prompts



```
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. 2023)



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.

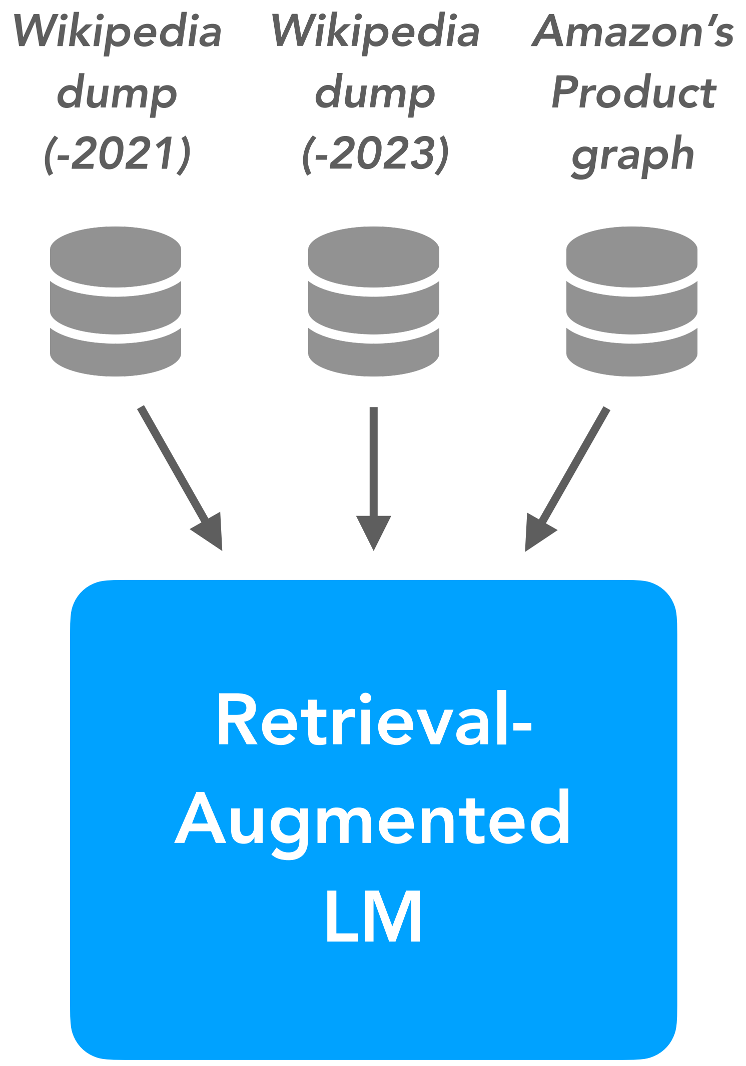
The Brown Act is California's law [WikiSearch("Brown Act") → The Ralph M. Brown Act is an act of the California State Legislature that guarantees the public's right to attend and participate in meetings of local legislative bodies.] that requires legislative bodies, like city councils, to hold their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

Toolformer (Schick et al. 2023)

Additional benefits of Augmented LMs

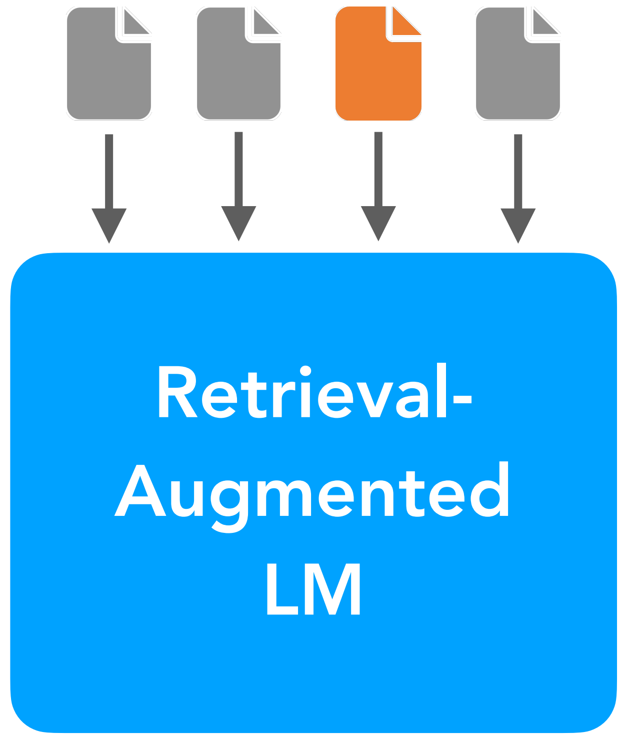
Modularity

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



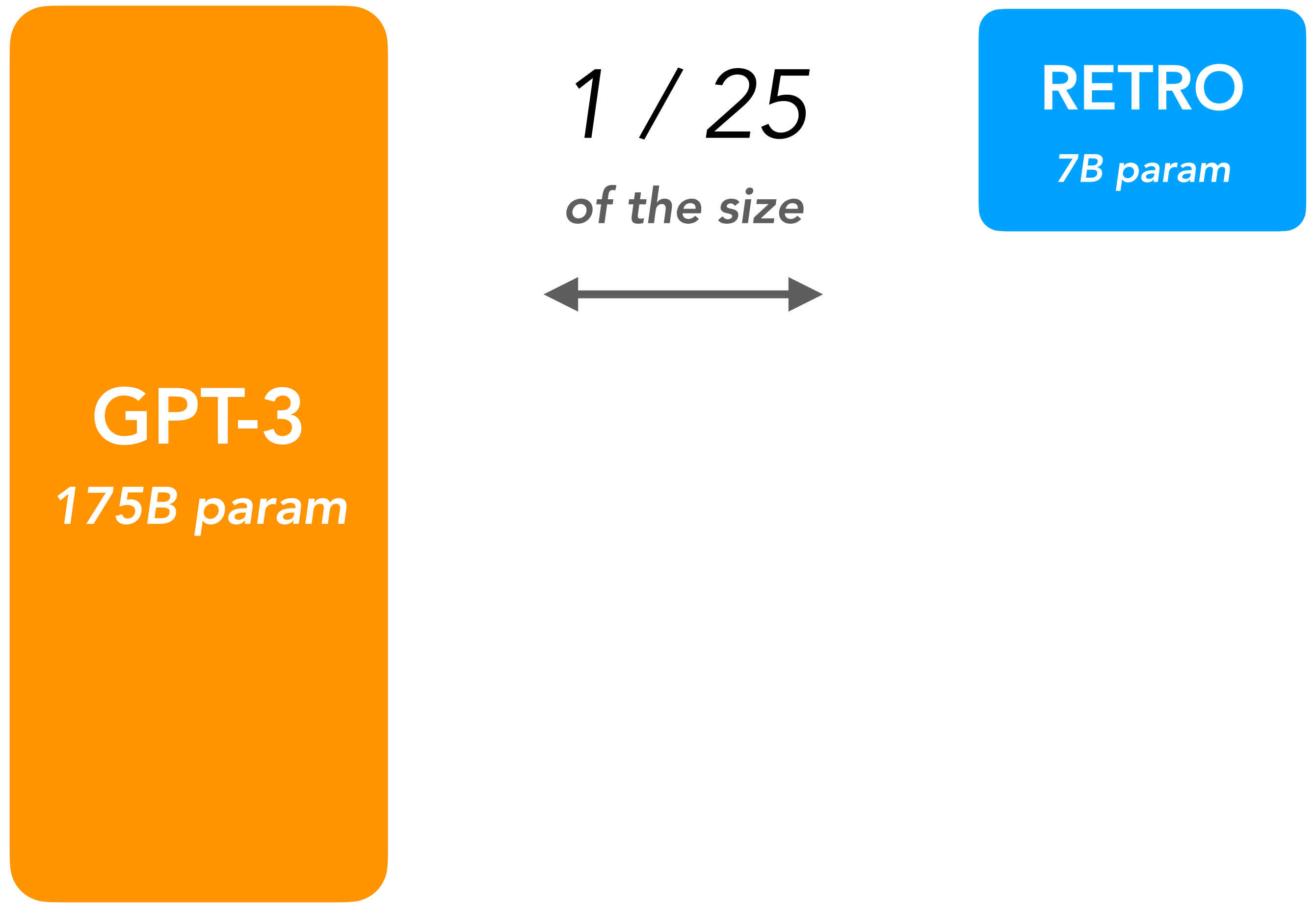
Explainability

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.



Recap

- **Retrieval-Augmented language models:**
  - Let us 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.
- **In the LLMs era**:
  - Retrieval aims to augment the prompt.
  - Models are interacting with various tools and APIs to enhance their reasoning capabilities.
- Using external knowledge can reduce the memorization stored in language model parameters and therefore reduce their size without compromising performance.

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).
- Borgeaud, Sebastian, et al. "Improving language models by retrieving from trillions of tokens." International conference on machine learning. PMLR, 2022.
- Mialon, Grégoire, et al. "Augmented language models: a survey." arXiv preprint arXiv:2302.07842 (2023).