

Question Answering & Reading Comprehension

Antoine Bosselut

EPFL



Announcements

- **No Lecture Tomorrow! Work on your project!**
 - Assignment 2 Grading Session in CE 1 6 at 1:15 PM — Last opportunity to discuss A2 grades.
- **Course Project:** Milestone 2 due **Sunday, May 26th!**
 - Feedback on milestone proposals released last week!
 - Student-collected data released last week!

Summer Internship Opportunities

- **Training Multilingual LLMs**
- **Diverse Sub-project Topics:**
 - Language Identification
 - Multilingual Data Collection
 - Multilingual Tokenization
 - Multilingual LLM architectures
- **Application Process:**
 - Send an e-mail to: nlp-projects-apply@groupes.epfl.ch
 - **Attach your CV and transcript and include [Summer Project] in your subject heading.**
- **Looking forward to collaborating with you!**

Fall Project Opportunity

- **Create a large language model** for a platform for BA1 courses, which includes a homemade "Moodle-like" system, with lecture notes, exercises and forum all in one place (see botafogo.saitis.net/analyse-1-GM).
- **Build a chatbot using open-source models** that can provide first answers to any kind of question that is asked about something on the platform, be it on the organisation, the lecture notes or the exercises.
- The model should be specifically trained on course teaching material, using in particular the 1500+ questions and answers already stored in the database.
- **Opportunity for Fall Semester 2024. Contact sacha.friedli@epfl.ch**

Today's Outline

- **Lecture**
 - **Question Answering:** Tasks, Models, Limitations
 - **Infusing non-parametric knowledge:** Retrieval-Augmented Language Models
 - ▶ How can we update the model's knowledge without updating its parameters?

Question Answering

**The goal of question answering is to build
systems that automatically answer questions
posed by humans in a natural language**

Question Answering Systems

Chatbots

AN

Can you give me a two-day itinerary for a weekend trip to Paris ? Please don't include more than three activities per day.

AI

Day 1:

1. Visit the Eiffel Tower - Start your day with the iconic symbol of Paris, the Eiffel Tower. Take a stroll around the Champ de Mars and enjoy the views from the tower.
2. Explore the Louvre Museum - After lunch, head to the Louvre Museum and admire the world-famous artwork such as the Mona Lisa and the Venus de Milo.
3. Stroll along the Seine River - In the evening, take a romantic stroll along the Seine River and enjoy the Parisian atmosphere.

Day 2:

1. Visit Notre-Dame Cathedral - Start your day by visiting one of the most famous Gothic cathedrals in the world, Notre-Dame Cathedral.
2. Explore the Marais neighborhood - After lunch, explore the Marais neighborhood, known for its historical architecture, trendy boutiques, and restaurants.
3. Enjoy a Cabaret Show - In the evening, enjoy a cabaret show at the famous Parisian venue, the Moulin Rouge, and end your trip with a memorable experience.

Question Answering Systems

Chatbots

AN
Can you give me a two-day itinerary for a weekend trip to Paris ? Please don't include more than three activities per day.

Day 1:

1. Visit the Eiffel Tower - Start your day with the iconic symbol of Paris, the Eiffel Tower. Take a stroll around the Champ de Mars and enjoy the views from the tower.
2. Explore the Louvre Museum - After lunch, head to the Louvre Museum and admire the world-famous artwork such as the Mona Lisa and the Venus de Milo.
3. Stroll along the Seine River - In the evening, take a romantic stroll along the Seine River and enjoy the Parisian atmosphere.

Day 2:

1. Visit Notre-Dame Cathedral - Start your day by visiting one of the most famous Gothic cathedrals in the world, Notre-Dame Cathedral.
2. Explore the Marais neighborhood - After lunch, explore the Marais neighborhood, known for its historical architecture, trendy boutiques, and restaurants.
3. Enjoy a Cabaret Show - In the evening, enjoy a cabaret show at the famous Parisian venue, the Moulin Rouge, and end your trip with a memorable experience.

Search Engines

who is the president of switzerland ?

All Images News Maps Videos More Tools

About 415'000'000 results (0.82 seconds)

Guy Parmelin

The President of the Swiss Confederation in 2021 is **Guy Parmelin** from the canton of Vaud. He was elected on 9 December 2020. The President's department in 2021 is the Federal Department of Economic Affairs, Education and Research EAER.

<https://www.admin.ch/gov/start/federal-presidency> :: Presidential year 2021

About featured snippets · Feedback

Question Answering Systems

Chatbots

AN
Can you give me a two-day itinerary for a weekend trip to Paris ? Please don't include more than three activities per day.

Day 1:

1. Visit the Eiffel Tower - Start your day with the iconic symbol of Paris, the Eiffel Tower. Take a stroll around the Champ de Mars and enjoy the views from the tower.
2. Explore the Louvre Museum - After lunch, head to the Louvre Museum and admire the world-famous artwork such as the Mona Lisa and the Venus de Milo.
3. Stroll along the Seine River - In the evening, take a romantic stroll along the Seine River and enjoy the Parisian atmosphere.

Day 2:

1. Visit Notre-Dame Cathedral - Start your day by visiting one of the most famous Gothic cathedrals in the world, Notre-Dame Cathedral.
2. Explore the Marais neighborhood - After lunch, explore the Marais neighborhood, known for its historical architecture, trendy boutiques, and restaurants.
3. Enjoy a Cabaret Show - In the evening, enjoy a cabaret show at the famous Parisian venue, the Moulin Rouge, and end your trip with a memorable experience.

Search Engines

who is the president of switzerland ?

All Images News Maps Videos More Tools

About 415'000'000 results (0.82 seconds)

Guy Parmelin

The President of the Swiss Confederation in 2021 is **Guy Parmelin** from the canton of Vaud. He was elected on 9 December 2020. The President's department in 2021 is the Federal Department of Economic Affairs, Education and Research EAER.

<https://www.admin.ch/gov/start/federal-presidency>

Forums / e-learning portals

ed CS-552 – Ed Discussion

top-p sampling: outputs similar to HF #275

Anonymous 2 weeks ago in Assignments ~ A3

Hello,
My top-p sampling implementation returns the same output as HF which seems a bit odd given the randomness factor of sampling. Is this a reason for concern?
Wishing you a lovely day and thank you for your guidance :D

Comment Edit Delete Endorse ...

PIN STAR WATCH 230 VIEWS

2 Answers

Deniz Bayazit STAFF 2 weeks ago

That's fine! It depends on how you modified the distribution before sampling. There are ways in which you can implement it slightly differently than HF, and the output will be different, but the implementation will be correct.
If it's the same, then congrats, you've replicated HF, and there is no problem :D

Comment Edit Delete Endorse ...

Components of QA

Input

Context

The Amazon rainforest, also known in English as Amazonia or the Amazon Jungle.

Question

Which name is also used to describe the Amazon rainforest in

Components of QA

Input

Context

The Amazon rainforest, also known in English as Amazonia or the Amazon Jungle.

Question

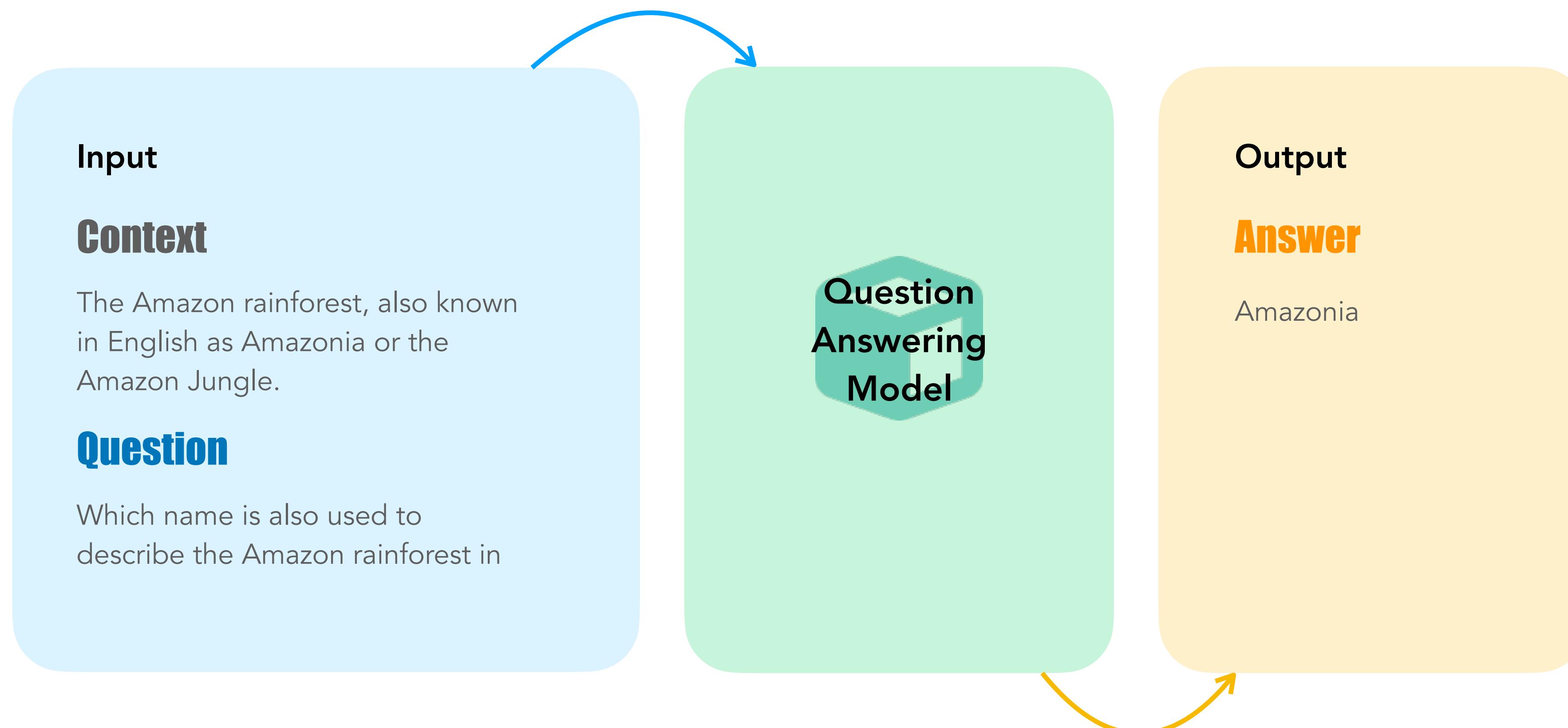
Which name is also used to describe the Amazon rainforest in

Output

Answer

Amazonia

Components of QA



How might we have historically designed a QA system?

Classical QA

(a) **CCG parse** builds an underspecified semantic representation of the sentence.

$$\begin{array}{cccc} \text{Former} & \text{municipalities} & \text{in} & \text{Brandenburg} \\ \hline N/N & N & N\backslash N/NP & NP \\ \lambda f \lambda x. f(x) \wedge \text{former}(x) & \lambda x. \text{municipalities}(x) & \lambda f \lambda x \lambda y. f(y) \wedge \text{in}(y, x) & \text{Brandenburg} \\ \hline N & & N\backslash N & \\ \lambda x. \text{former}(x) \wedge \text{municipalities}(x) & & \lambda f \lambda y. f(y) \wedge \text{in}(y, \text{Brandenburg}) & \\ \hline N & & & \\ l_0 = \lambda x. \text{former}(x) \wedge \text{municipalities}(x) \wedge \text{in}(x, \text{Brandenburg}) & & & \end{array}$$

(b) **Constant matches** replace underspecified constants with Freebase concepts

$$l_0 = \lambda x. \text{former}(x) \wedge \text{municipalities}(x) \wedge \text{in}(x, \text{Brandenburg})$$

$$l_1 = \lambda x. \text{former}(x) \wedge \text{municipalities}(x) \wedge \text{in}(x, \text{Brandenburg})$$

$$l_2 = \lambda x. \text{former}(x) \wedge \text{municipalities}(x) \wedge \text{location.containedby}(x, \text{Brandenburg})$$

$$l_3 = \lambda x. \text{former}(x) \wedge \text{OpenRel}(x, \text{Municipality}) \wedge \text{location.containedby}(x, \text{Brandenburg})$$

$$l_4 = \lambda x. \text{OpenType}(x) \wedge \text{OpenRel}(x, \text{Municipality}) \wedge \text{location.containedby}(x, \text{Brandenburg})$$

- Convert text to logical forms from text and execute against structured databases

What might be a challenge of this approach?

Complexity of QA

- Sources of information:
 - Text passages, knowledge bases, tables, images
- Question types:
 - Factoid vs. commonsense, open-domain vs. Close-domain, simple vs. multi-hop
- Answer type:
 - Short snippet, paragraph long answer, yes / no questions, numerical...

Classical QA

(a) **CCG parse** builds an underspecified semantic representation of the sentence.

$$\begin{array}{cccc} \text{Former} & \text{municipalities} & \text{in} & \text{Brandenburg} \\ \hline N/N & N & N\backslash N/NP & NP \\ \lambda f \lambda x. f(x) \wedge \text{former}(x) & \lambda x. \text{municipalities}(x) & \lambda f \lambda x \lambda y. f(y) \wedge \text{in}(y, x) & \text{Brandenburg} \\ \hline N & & N\backslash N & \\ \lambda x. \text{former}(x) \wedge \text{municipalities}(x) & & \lambda f \lambda y. f(y) \wedge \text{in}(y, \text{Brandenburg}) & \\ \hline N & & & \\ l_0 = \lambda x. \text{former}(x) \wedge \text{municipalities}(x) \wedge \text{in}(x, \text{Brandenburg}) & & & \end{array}$$

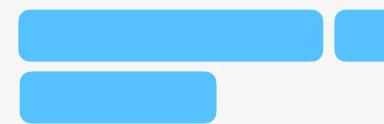
(b) **Constant matches** replace underspecified constants with Freebase concepts

$$\begin{aligned} l_0 &= \lambda x. \text{former}(x) \wedge \text{municipalities}(x) \wedge \text{in}(x, \text{Brandenburg}) \\ l_1 &= \lambda x. \text{former}(x) \wedge \text{municipalities}(x) \wedge \text{in}(x, \text{Brandenburg}) \\ l_2 &= \lambda x. \text{former}(x) \wedge \text{municipalities}(x) \wedge \text{location.containedby}(x, \text{Brandenburg}) \\ l_3 &= \lambda x. \text{former}(x) \wedge \text{OpenRel}(x, \text{Municipality}) \wedge \text{location.containedby}(x, \text{Brandenburg}) \\ l_4 &= \lambda x. \text{OpenType}(x) \wedge \text{OpenRel}(x, \text{Municipality}) \wedge \text{location.containedby}(x, \text{Brandenburg}) \end{aligned}$$

- Convert text to logical forms from text and execute against structured databases
- **Challenge:** Dealing with open-domain data and relationships outside DB

Types of QA

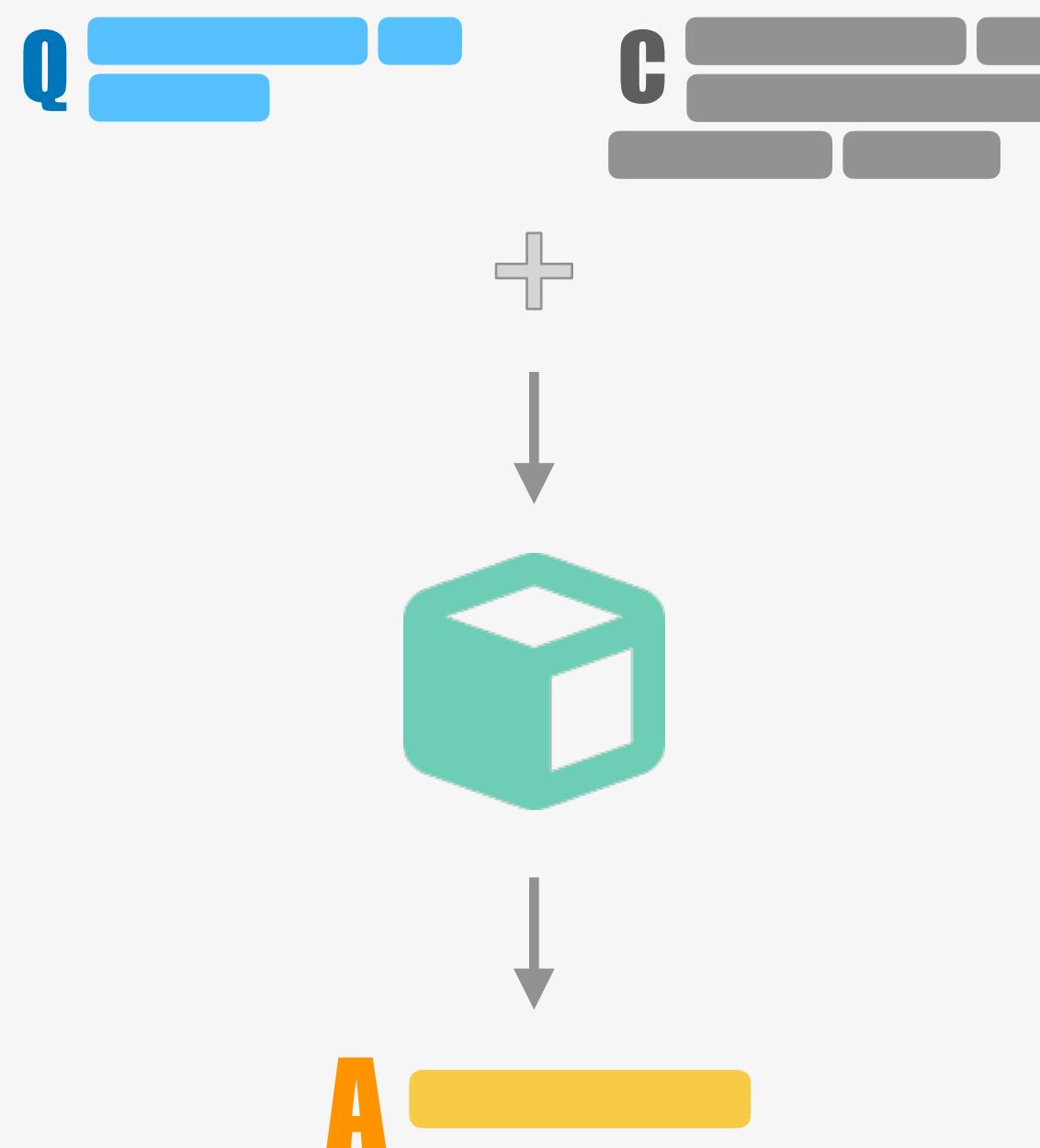
Extractive QA

Q 

C 

Types of QA

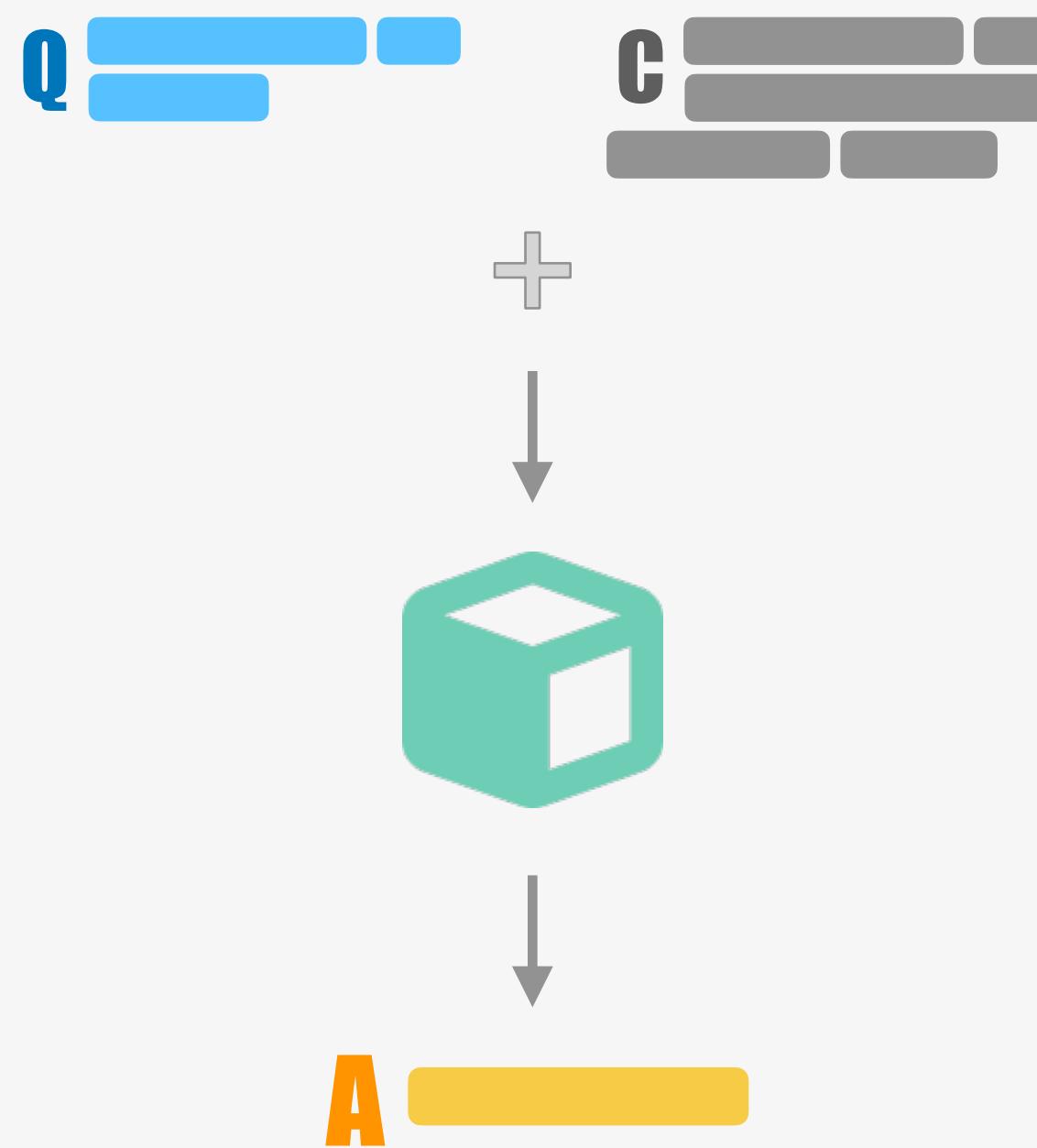
Extractive QA



Answer is extracted
from the context

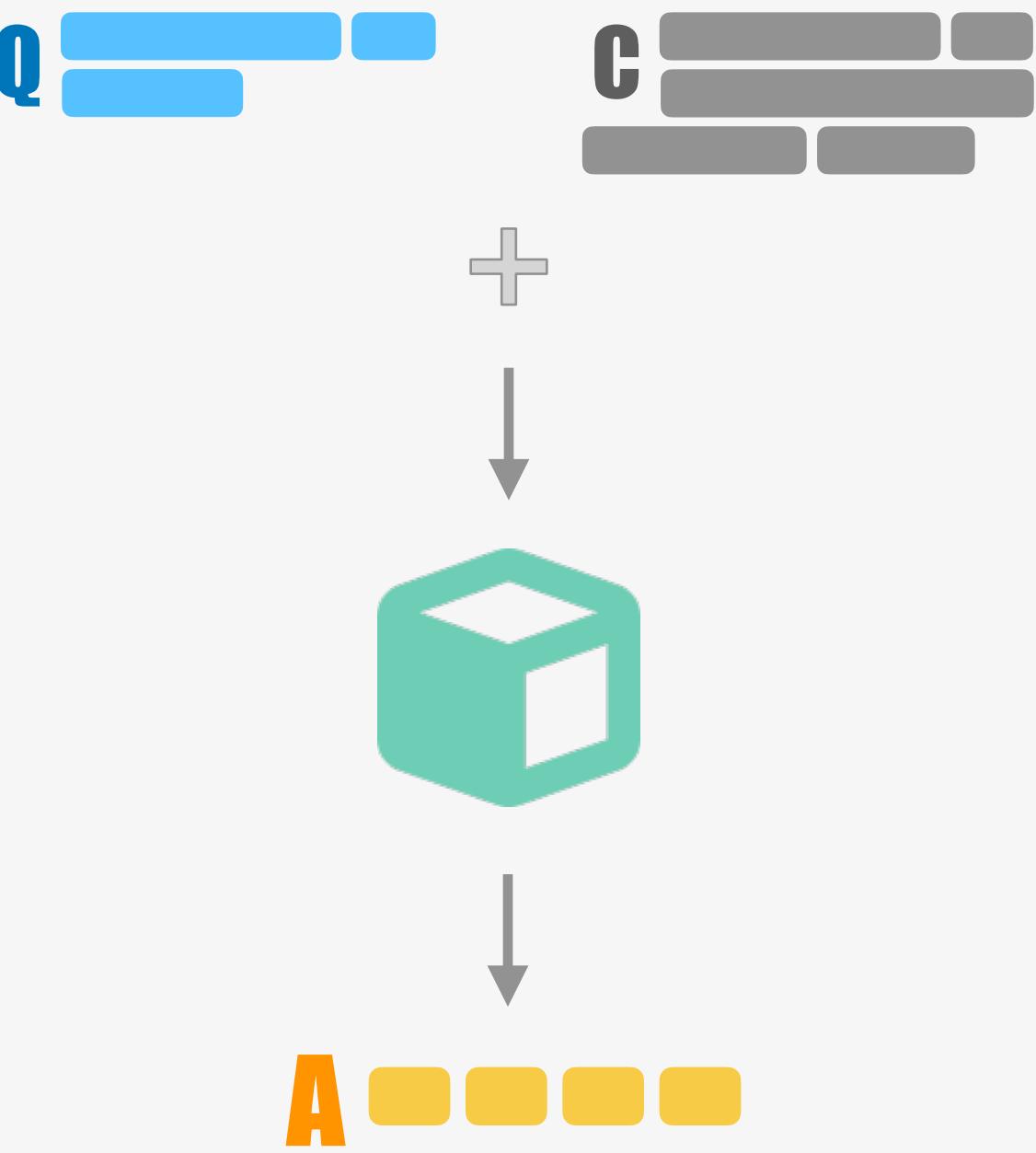
Types of QA

Extractive QA



Answer is extracted
from the context

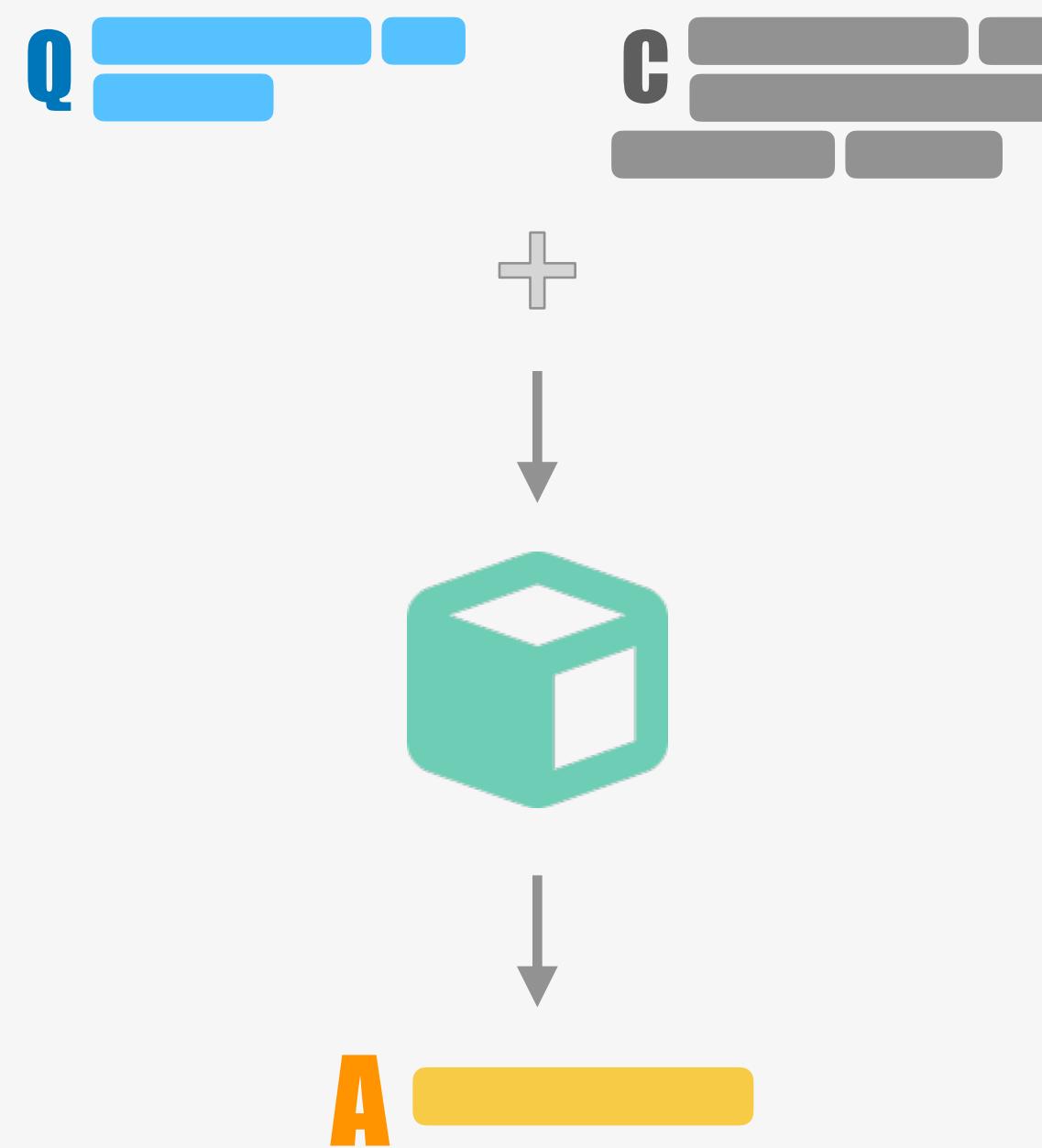
Open-Generative QA



Answer is generated
in an auto-regressive way

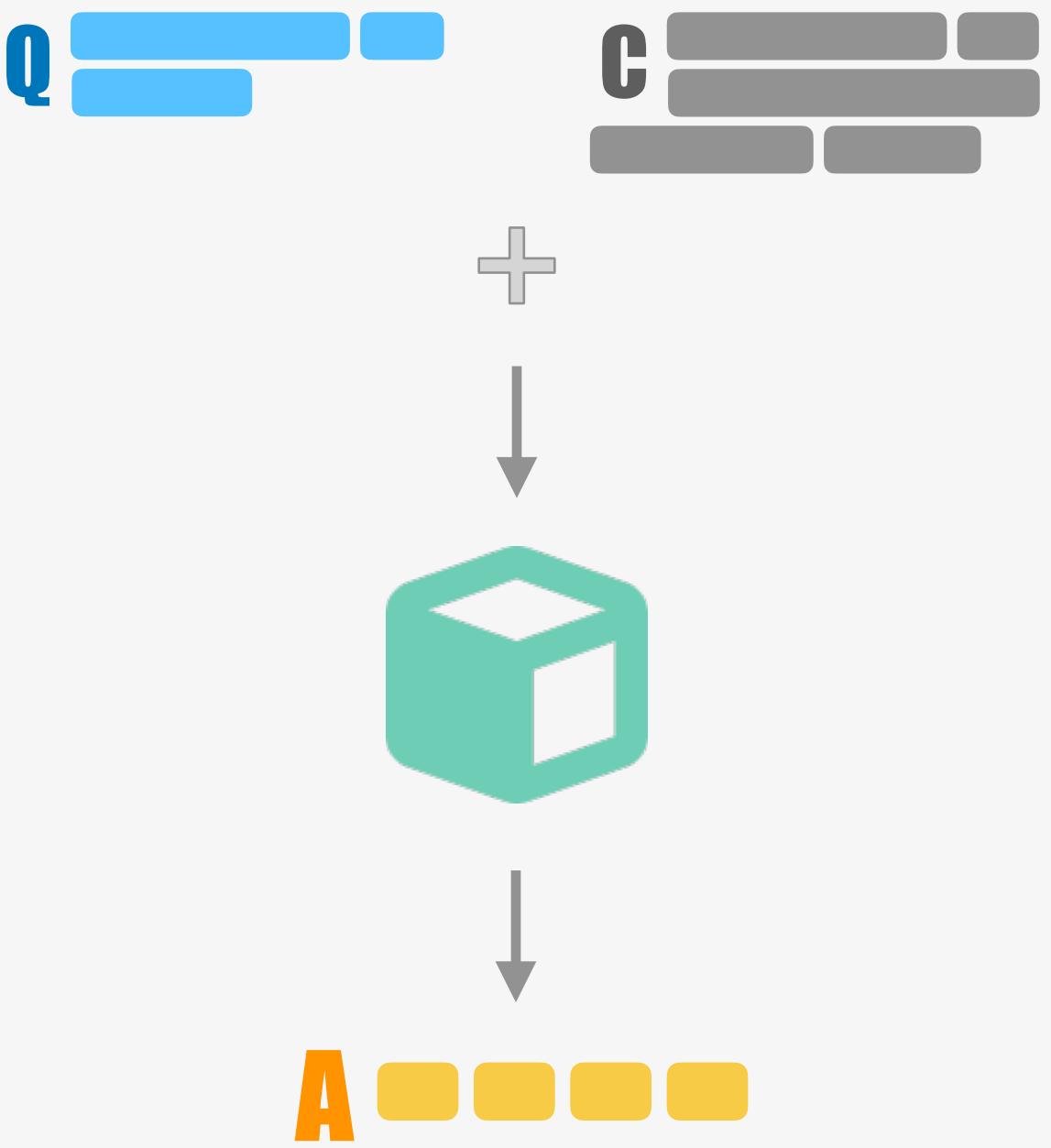
Types of QA

Extractive QA



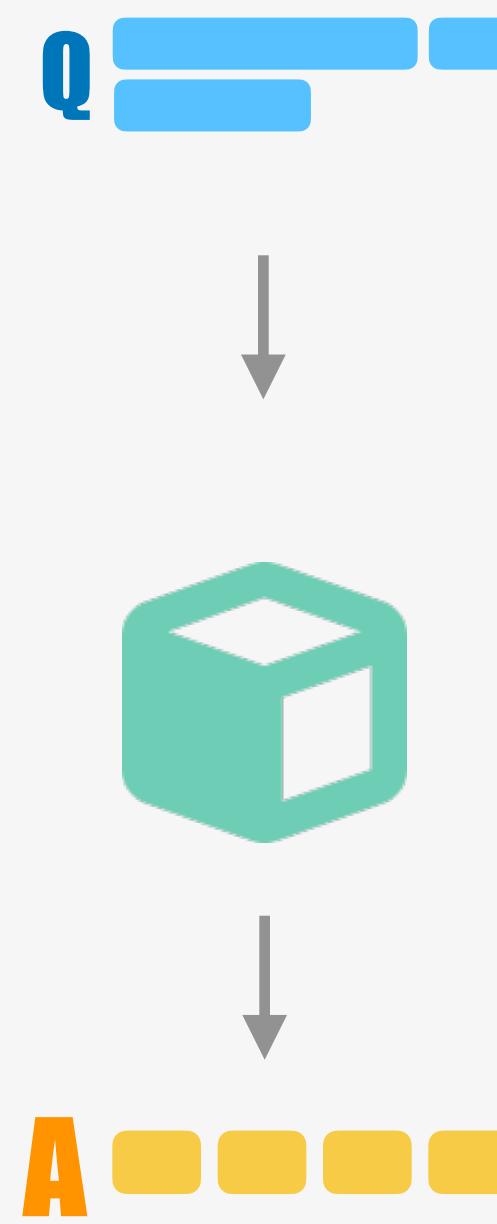
Answer is extracted
from the context

Open-Generative QA



Answer is generated
in an auto-regressive way

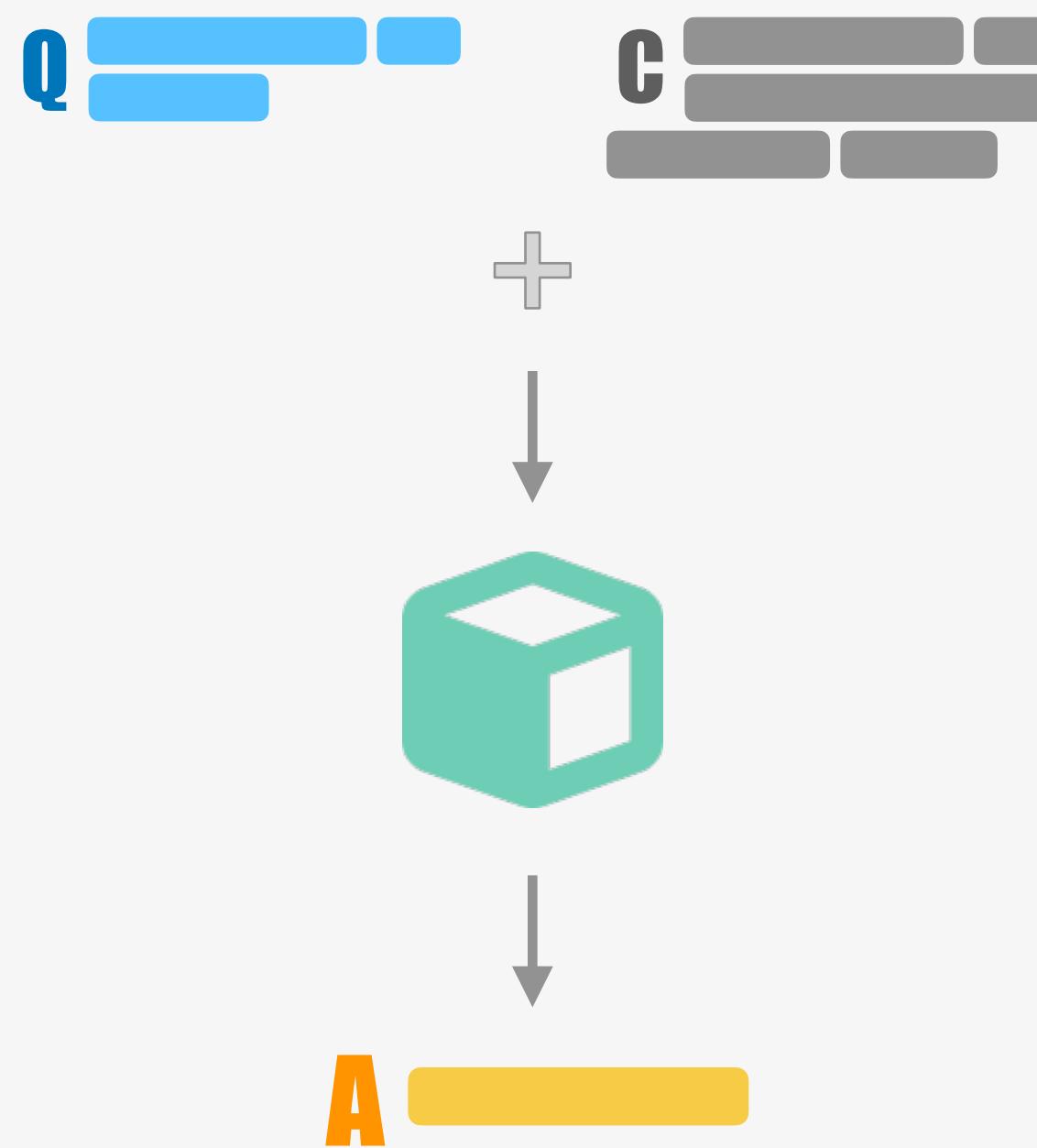
Closed-Generative QA



Answer is generated
in an auto-regressive way

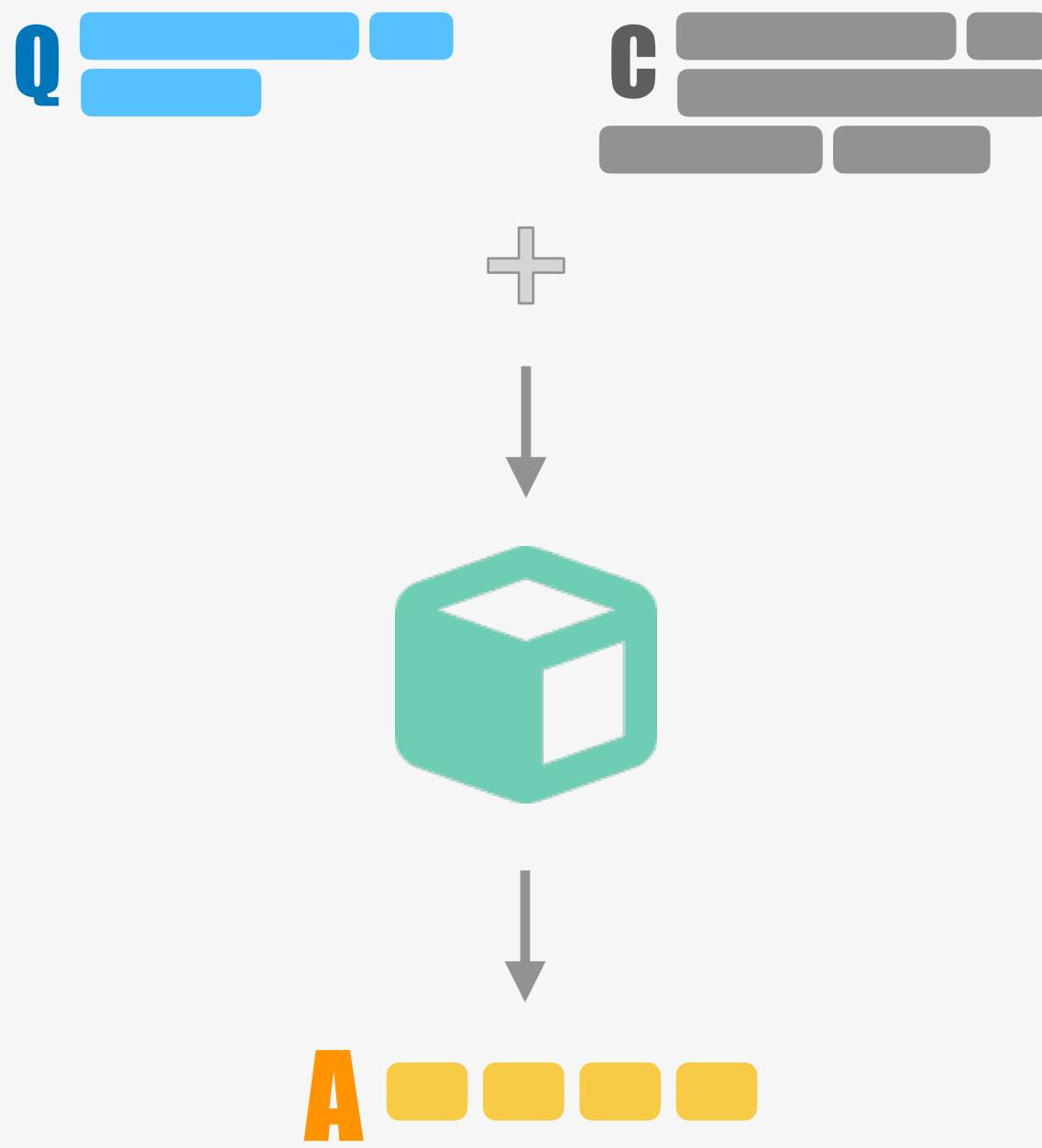
Types of QA

Extractive QA



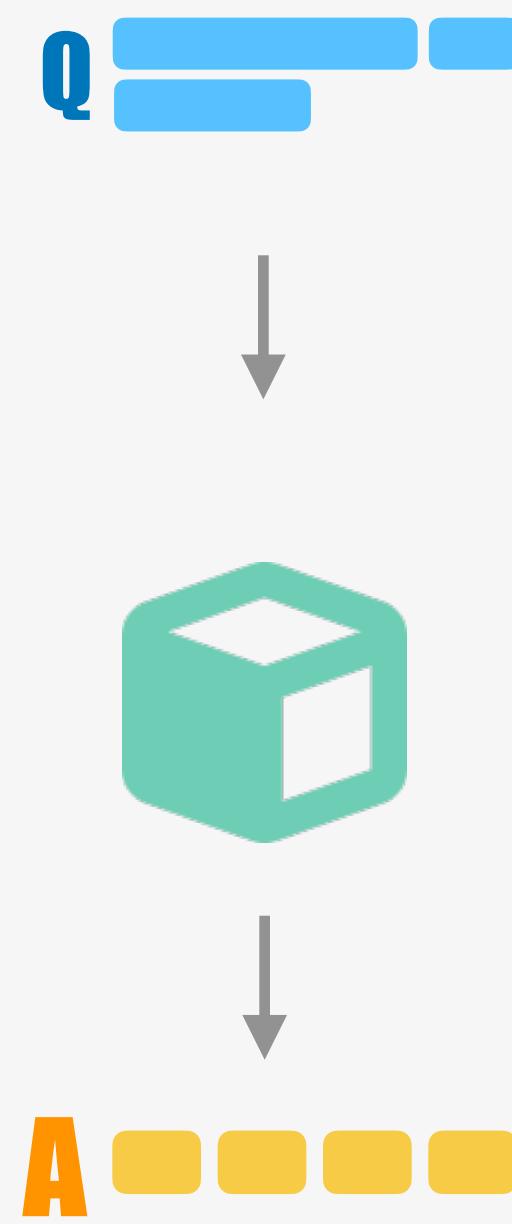
Answer is extracted
from the context

Open-Generative QA



Answer is generated
in an auto-regressive way

Closed-Generative QA



Answer is generated
in an auto-regressive way

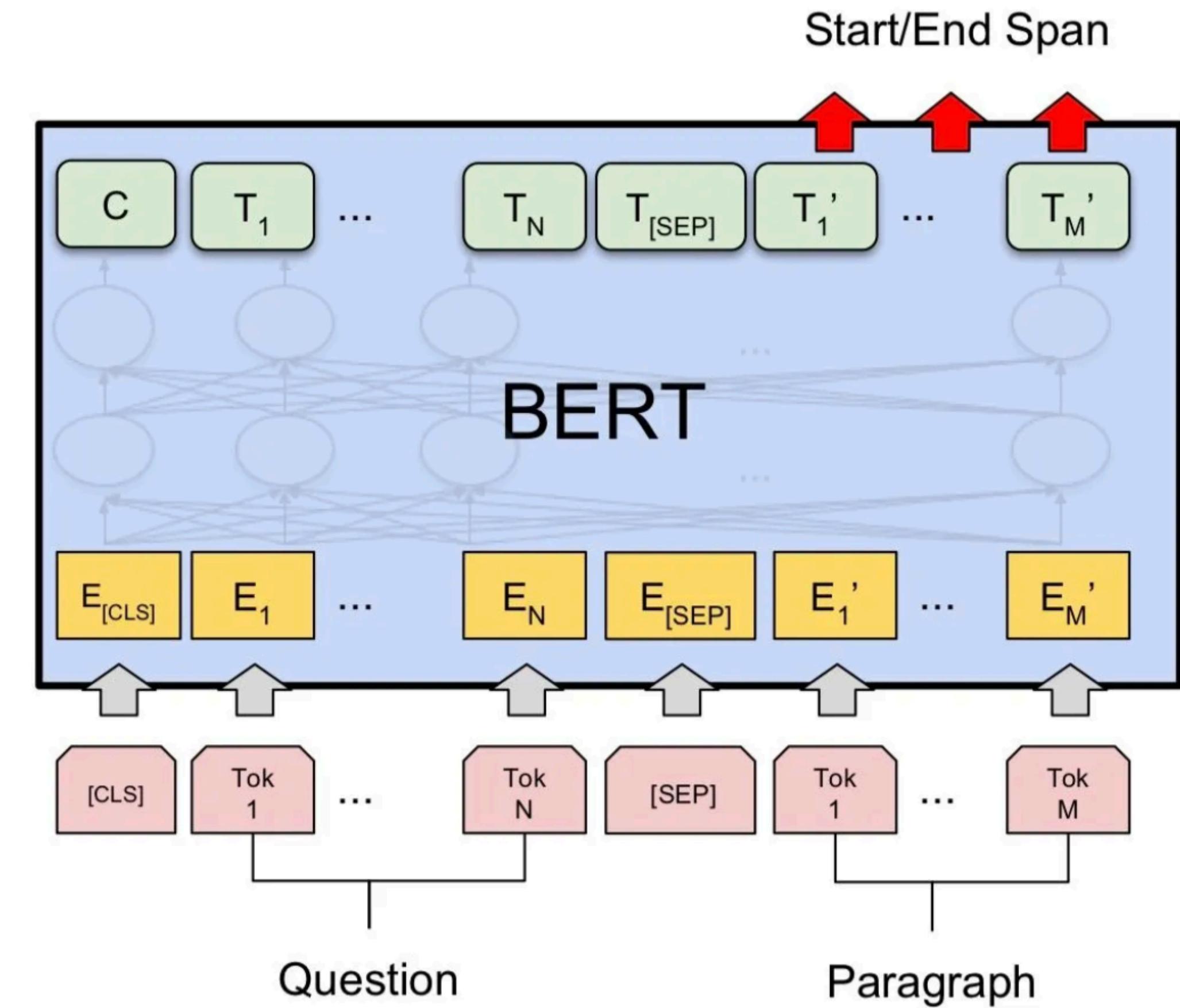
Open Book
Context is available

Closed Book

Extractive QA

Goal:

Predict the **start** and **end** tokens of the answer in the context.



Extractive QA

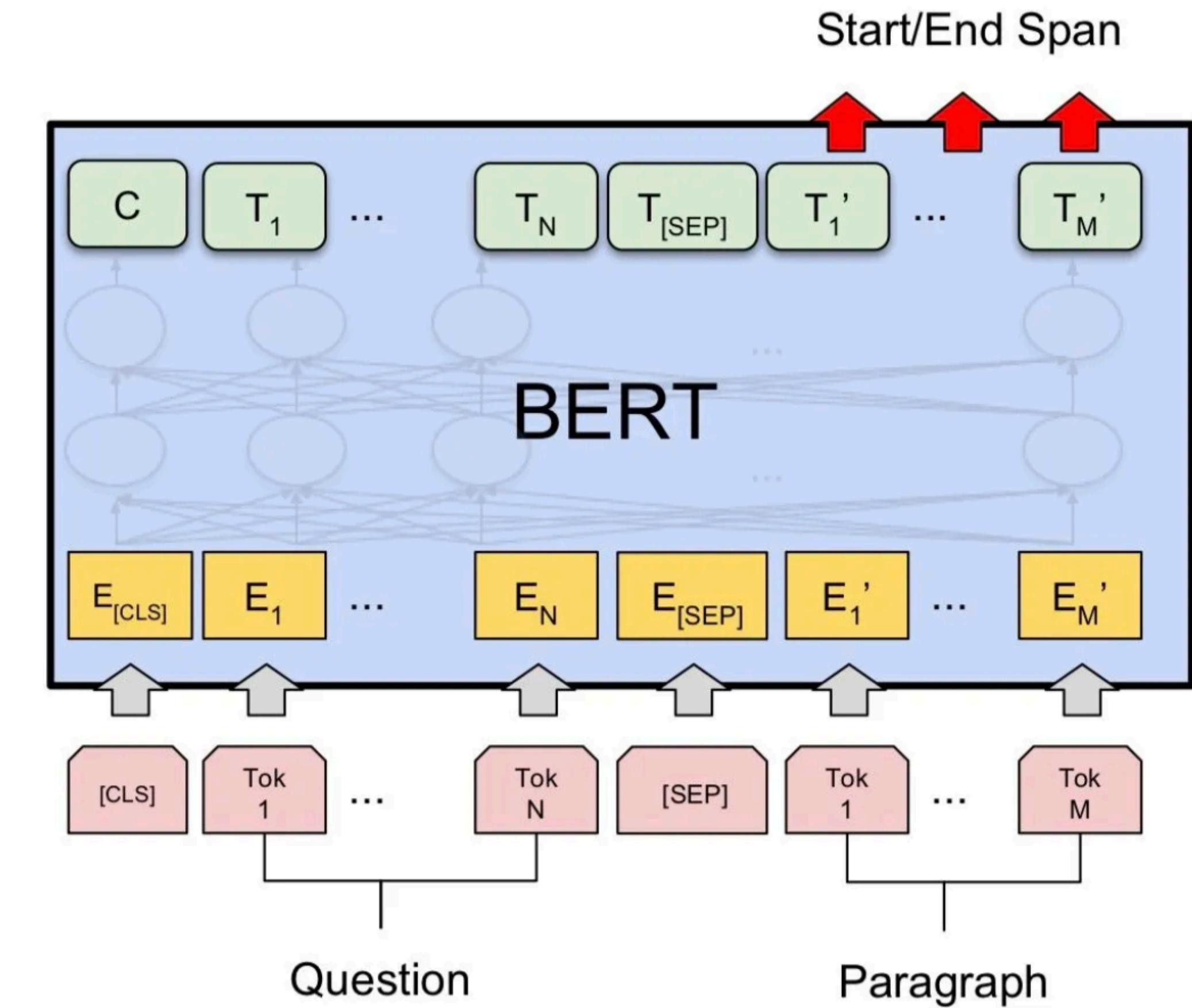
Goal:

Predict the **start** and **end** tokens of the answer in the context.

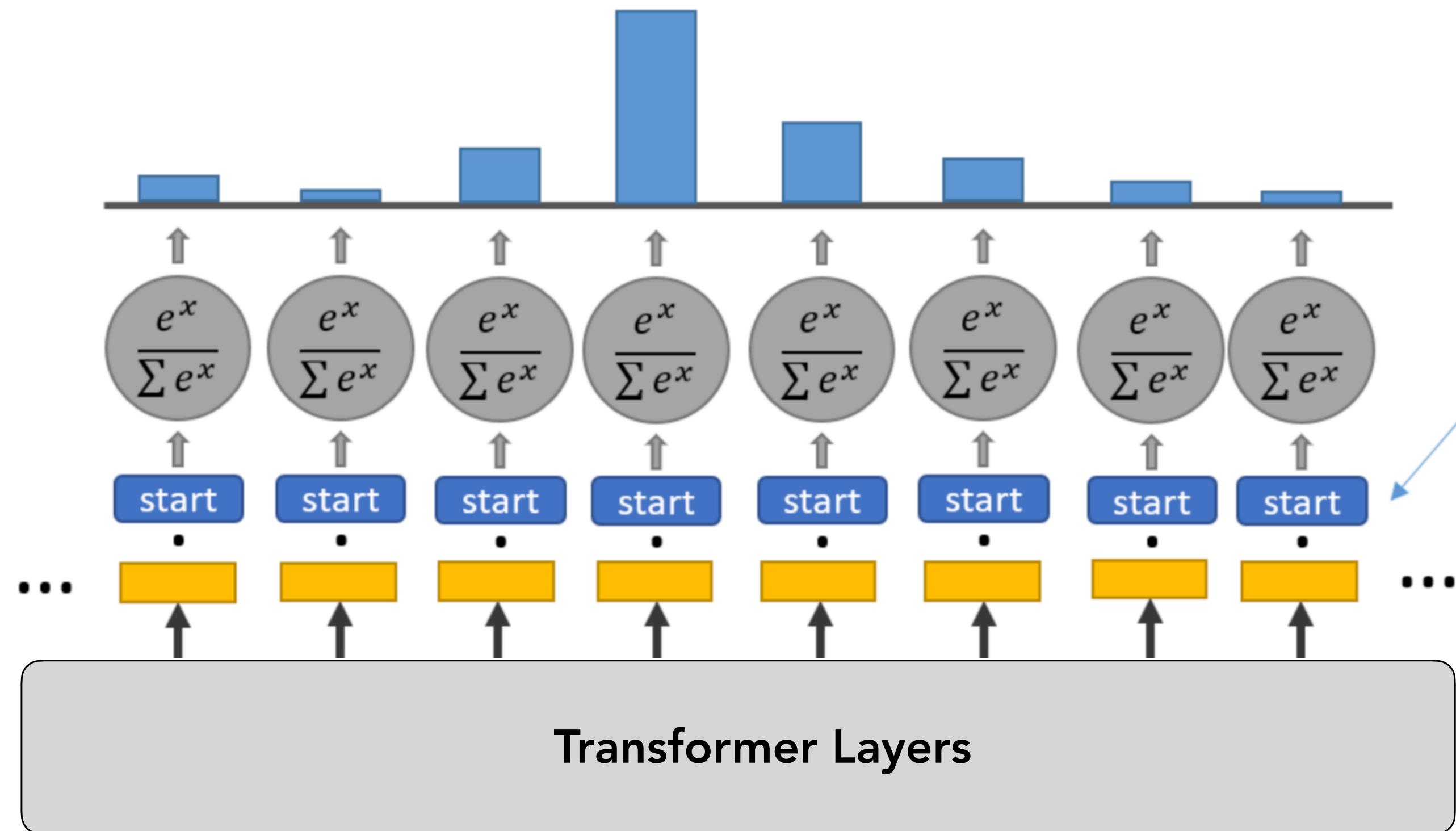
Input:

- The models are a function of the **question** and the **context** together.

<question> <SEP> <context>



Extractive QA



- We add 2 linear layers: one for the **start** position & another for the **end** position.
- We have separate weights for each of them. During training, they are trained together.
- After taking the dot product between the output embeddings and the **start linear layer** weights, we apply the softmax activation to produce a probability distribution over all of the words.

The token with the highest probability is selected as the start token.

QA Datasets: SQuaD

Context

The Normans (Norman: Nourmands; French: Normands; Latin: Normanni) were the people who in the 10th and 11th centuries gave their name to Normandy, a region in France. They were descended from Norse ("Norman" comes from "Norseman") raiders and pirates from Denmark, Iceland and Norway [...]

Question

When were the Normans in Normandy?



Answers

- 10th and 11th centuries
- In the 10th and 11th centuries

QA Datasets: SQuaD

Context

The Normans (Norman: Nourmands; French: Normands; Latin: Normanni) were the people who in the 10th and 11th centuries gave their name to Normandy, a region in France. They were descended from Norse ("Norman" comes from "Norseman") raiders and pirates from Denmark, Iceland and Norway [...]

Train split	130K
Test split	12K
Unanswerable	50K

Question

When were the Normans in Normandy?



Answers

- 10th and 11th centuries
- In the 10th and 11th centuries

QA Datasets: SQuAD

Context

The Normans (Norman: Nourmands; French: Normands; Latin: Normanni) were the people who in the 10th and 11th centuries gave their name to Normandy, a region in France. They were descended from Norse ("Norman" comes from "Norseman") raiders and pirates from Denmark, Iceland and Norway [...]

Train split	130K
Test split	12K
Unanswerable	50K

Question

When were the Normans in Normandy?

Answers



Answer

starts

- 10th and 11th centuries
- In the 10th and 11th centuries

94

87

QA Datasets: SQuaD

Context

The Normans (Norman: Nourmands; French: Normands; Latin: Normanni) were the people who in the 10th and 11th centuries gave their name to Normandy, a region in France. They were descended from Norse ("Norman" comes from "Norseman") raiders and pirates from Denmark, Iceland and Norway [...]

Train split	130K
Test split	12K
Unanswerable	50K

Question

When were the Normans in Normandy?

Required reasoning

- **Cross-sentence:** The partial answer can be located in multiple sentences.
- **Lexical and syntactic variations:** Synonyms & paraphrasing
- **World knowledge:** The answer sentence also requires commonsense knowledge to resolve.

Answers

- 10th and 11th centuries
- In the 10th and 11th centuries



Answer
starts

94

87

Why is Extractive QA popular?

- Extractive QA is closed-form task
 - No need to generate open-world answers (only need to highlight spans)
- SQuAD was **big**:
 - > 100K questions when data-driven deep learning was exploding (e.g., LSTMs)
- Progress on dataset was easy to make
 - Lots of people wanted to work on it and large improvement could be made over classical methods

Saturation

SQuAD1.1 Leaderboard

Here are the ExactMatch (EM) and F1 scores evaluated on the test set of SQuAD v1.1.

Rank	Model	EM	F1
	Human Performance <i>Stanford University</i> (Rajpurkar et al. '16)	82.304	91.221
1	{ANNA} (single model) <i>LG AI Research</i>	90.622	95.719
2	LUKE (single model) <i>Studio Ousia & NAIST & RIKEN AIP</i> https://arxiv.org/abs/2010.01057	90.202	95.379
3	XLNet (single model) <i>Google Brain & CMU</i>	89.898	95.080
4	XLNET-123++ (single model) <i>MST/EOI</i> http://tia.today	89.856	94.903
4	XLNET-123 (single model) <i>MST/EOI</i>	89.646	94.930
5	SpanBERT (single model) <i>FAIR & UW</i>	88.839	94.635

Leaderboard

SQuAD2.0 tests the ability of a system to not only answer reading comprehension questions, but also abstain when presented with a question that cannot be answered based on the provided paragraph.

Rank	Model	EM	F1
	Human Performance <i>Stanford University</i> (Rajpurkar & Jia et al. '18)	86.831	89.452
1	IE-Net (ensemble) <i>RICOH_SRCB_DML</i>	90.939	93.214
2	FPNet (ensemble) <i>Ant Service Intelligence Team</i>	90.871	93.183
3	IE-NetV2 (ensemble) <i>RICOH_SRCB_DML</i>	90.860	93.100
4	SA-Net on Albert (ensemble) <i>QIANXIN</i>	90.724	93.011
5	SA-Net-V2 (ensemble) <i>QIANXIN</i>	90.679	92.948
5	Retro-Reader (ensemble) <i>Shanghai Jiao Tong University</i> http://arxiv.org/abs/2001.09694	90.578	92.978
5	FPNet (ensemble) <i>YuYang</i>	90.600	92.899

Is Reading Comprehension Solved?

Article: Super Bowl 50

Paragraph: “*Peyton Manning became the first quarterback ever to lead two different teams to multiple Super Bowls. He is also the oldest quarterback ever to play in a Super Bowl at age 39. The past record was held by John Elway, who led the Broncos to victory in Super Bowl XXXIII at age 38 and is currently Denver’s Executive Vice President of Football Operations and General Manager. Quarterback Jeff Dean had jersey number 37 in Champ Bowl XXXIV.*”

Question: “*What is the name of the quarterback who was 38 in Super Bowl XXXIII?*”

Original Prediction: John Elway

Prediction under adversary: Jeff Dean

	Match Single	Match Ens.	BiDAF Single	BiDAF Ens.
Original	71.4	75.4	75.5	80.0
ADDSENT	27.3	29.4	34.3	34.2
ADDONESENT	39.0	41.8	45.7	46.9
ADDANY	7.6	11.7	4.8	2.7
ADDCOMMON	38.9	51.0	41.7	52.6

Systems perform much worse on adversarial samples with distractor information

Generative QA

- Generative models output the answer one token at a time.
- For both Open-Book (with context) and Closed-Book (without context) we can use Autoregressive LMs (**GPT** variants) or Sequence-to-Sequence models (**T5**, **BART**).
- Models are fine-tuned for the Question Answering task by being presented with multiple question-answer choices across numerous examples.

Generative QA

- Generative models output the answer one token at a time.
- For both Open-Book (with context) and Closed-Book (without context) we can use Autoregressive LMs (**GPT** variants) or Sequence-to-Sequence models (**T5, BART**).
- Models are fine-tuned for the Question Answering task by being presented with multiple question-answer choices across numerous examples.

Input: question **q** & context **c**
Output: probability of an answer **a** based on model parameters **θ**

$$P(a|c, q; \theta) = \prod_{i=1}^{|a|} P(a_i|c, q, a_{<i}; \theta)$$

Generative QA

- Generative models output the answer one token at a time.
- For both Open-Book (with context) and Closed-Book (without context) we can use Autoregressive LMs (**GPT** variants) or Sequence-to-Sequence models (**T5, BART**).
- Models are fine-tuned for the Question Answering task by being presented with multiple question-answer choices across numerous examples.

Input: question \mathbf{q} & context \mathbf{c}
Output: probability of an answer \mathbf{a} based on model parameters θ

$$P(\mathbf{a}|\mathbf{c}, \mathbf{q}; \theta) = \prod_{i=1}^{|\mathbf{a}|} P(a_i|\mathbf{c}, \mathbf{q}, \mathbf{a}_{<i}; \theta)$$

Many generative QA datasets

Most other tasks can be framed as a generative QA task

Extractive vs Generative QA

Pros of Extractive:

- Syntactic and Lexical consistency
- Factual accuracy

Extractive vs Generative QA

Pros of Extractive:

- Syntactic and Lexical consistency
- Factual accuracy

Cons of Extractive:

- Cannot generate unique/novel utterances
- Rigid output (Truncated sentences, etc.)

Extractive vs Generative QA

Pros of Extractive:

- Syntactic and Lexical consistency
- Factual accuracy

Pros of Generative:

- Human-like response structures
- Suited for long-form answers
- Suited for cross-sentence reasoning

Cons of Extractive:

- Cannot generate unique/novel utterances
- Rigid output (Truncated sentences, etc.)

Extractive vs Generative QA

Pros of Extractive:

- Syntactic and Lexical consistency
- Factual accuracy

Pros of Generative:

- Human-like response structures
- Suited for long-form answers
- Suited for cross-sentence reasoning

Cons of Extractive:

- Cannot generate unique/novel utterances
- Rigid output (Truncated sentences, etc.)

Cons of Generative:

- Hallucinations / lexical repetitions
- Grammar mistakes

Extractive vs Generative QA

Pros of Extractive:

- Syntactic and Lexical consistency
- Factual accuracy

Pros of Generative:

- Human-like response structures
- Suited for long-form answers
- Suited for cross-sentence reasoning

Cons of Extractive:

- Cannot generate unique/novel utterances
- Rigid output (Truncated sentences, etc.)

Cons of Generative:

- Hallucinations / lexical repetitions
- Grammar mistakes

The choice of the QA system depends highly
on user requirements and its application.

How should we evaluate QA systems?

Evaluation of QA systems

Exact match (EM)

Percentage of predictions that match any one of the ground truth answers exactly.

```
1 if str(golden_answer) == str(pred_answer) else 0
```

Evaluation of QA systems

Exact match (EM)

Percentage of predictions that match any one of the ground truth answers exactly.

```
1 if str(golden_answer) == str(pred_answer) else 0
```

"Who is the president of France?"

Golden answer	Predicted answers	EM
Emmanuel Macron	Emmanuel Macron	✓
Emmanuel Jean-Michel Frédéric Macron		✗

Evaluation of QA systems

Exact match (EM)

Percentage of predictions that match any one of the ground truth answers exactly.

```
1 if str(golden_answer) == str(pred_answer) else 0
```

F1 score

Measures the average token overlap between the prediction and ground truth answer.

(more forgiving than EM)

"Who is the president of France?"

Golden answer	Predicted answers	EM	F1
Emmanuel Macron	Emmanuel Macron	✓	✓
	Emmanuel Jean-Michel Frédéric Macron	✗	✓

Evaluation of QA systems

Exact match (EM)

Percentage of predictions that match any one of the ground truth answers exactly.

```
1 if str(golden_answer) == str(pred_answer) else 0
```

F1 score

Measures the average token overlap between the prediction and ground truth answer.

(more forgiving than EM)

Complimentary approaches

- **Top-k:** Compute EM or F1 score after extracting/generating top-k answers
- **Post-process output:** Lemmatize answers, remove stop words, etc. before computing EM & F1 scores.

"Who is the president of France?"

Golden answer	Predicted answers	EM	F1
Emmanuel Macron	Emmanuel Macron	✓	✓
	Emmanuel Jean-Michel Frédéric Macron	✗	✓

Evaluation of QA systems

Exact match (EM)

Percentage of predictions that match any one of the ground truth answers exactly.

```
1 if str(golden_answer) == str(pred_answer) else 0
```

F1 score

Measures the average token overlap between the prediction and ground truth answer.

(more forgiving than EM)

Complimentary approaches

- **Top-k:** Compute EM or F1 score after extracting/generating top-k answers
- **Post-process output:** Lemmatize answers, remove stop words, etc. before computing EM & F1 scores.

"Who is the president of France?"

Golden answer	Predicted answers	EM	F1
Emmanuel Macron	Emmanuel Macron	✓	✓
	Emmanuel Jean-Michel Frédéric Macron	✗	✓

Challenge:
Semantic answer similarity

Two answers can be equivalent even if they don't share the same tokens.

GOLDEN
100%



PRED
One hundred percent

What about the evaluation of long-form answers?

Long Form QA - Evaluation

Natural Questions dataset

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

Qualitative measures

- Topical
- Fluent
- Coherent
- Commonsense
- Etc.

Long Form QA - Evaluation

Natural Questions dataset

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

Qualitative measures

- Topical
- Fluent
- Coherent
- Commonsense
- Etc.

Quantitative measures

Similar to text generation evaluation metrics

- Content overlap metrics (ROUGE, BLEU, etc.)
- Model-based metrics (BERTScore etc.)

What do QA systems look like today?

QA in LLM era

MAIN IDEA:

**Customization of the prompt to answer questions
with different output structure by providing in-
context demonstrations (i.e., few-shot exemplars).**

QA in LLM era

Information-retrieval

Context: The 2007–2008 financial crisis, or Global Financial Crisis (GFC), was a severe worldwide economic crisis that occurred in the early 21st century. [...]

Question: What caused the financial crisis in 2008?

Answer:

- Housing bubble
- Borrowers unable to pay their loans

MAIN IDEA:

Customization of the prompt to answer questions
with different output structure by providing in-
context demonstrations (i.e., few-shot exemplars).

QA in LLM era

Information-retrieval

Context: The 2007–2008 financial crisis, or Global Financial Crisis (GFC), was a severe worldwide economic crisis that occurred in the early 21st century. [...]

Question: What caused the financial crisis in 2008?

Answer:

- Housing bubble
- Borrowers unable to pay their loans

Graph-extraction

Context: The 2007–2008 financial crisis, or Global Financial Crisis (GFC), was a severe worldwide economic crisis that occurred in the early 21st century. [...]

Question: What caused what in the context above?

Answer:

```
| Cause | Effect |
| Housing bubble | 2008 Financial crisis |
|end|
```

MAIN IDEA:

Customization of the prompt to answer questions
with different output structure by providing in-
context demonstrations (i.e., few-shot exemplars).

QA in LLM era

Information-retrieval

Context: The 2007–2008 financial crisis, or Global Financial Crisis (GFC), was a severe worldwide economic crisis that occurred in the early 21st century. [...]

Question: What caused the financial crisis in 2008?

Answer:

- Housing bubble
- Borrowers unable to pay their loans

Graph-extraction

Context: The 2007–2008 financial crisis, or Global Financial Crisis (GFC), was a severe worldwide economic crisis that occurred in the early 21st century. [...]

Question: What caused what in the context above?

Answer:

```
| Cause | Effect |
| Housing bubble | 2008 Financial crisis |
|end|
```

Chain of Thought

Context: The 2007–2008 financial crisis, or Global Financial Crisis (GFC), was a severe worldwide economic crisis that occurred in the early 21st century. [...]

Question: Did the housing bubble cause the 2008 financial crisis?

Answer: Yes / No <reason> because . . .

MAIN IDEA:

Customization of the prompt to answer questions with different output structure by providing in-context demonstrations (i.e., few-shot exemplars).

What challenges remain?

Challenges & Limitations

Synonymity & Ambiguity

Syntactic, lexical or semantic divergence between the question and the context.

Question: Which governing bodies have veto power?

Context: The European Parliament and the Council of the European Union have powers of amendment and veto during the legislative process.

Challenges & Limitations

Synonymity & Ambiguity

Syntactic, lexical or semantic divergence between the question and the context.

Question: Which **governing bodies** have veto power?

Context: The European Parliament and the Council of the European Union have powers of amendment and veto during the legislative process.

Multi-hop reasoning

The answer might spread across different sentences, different documents, and different logical steps.

Question: Who is Florence for Betty ?

Context: Natasha is a granddaughter to **Betty**. **Florence** is Gregorio 's sister. Gregorio is a brother of Natasha.

Challenges & Limitations

Synonymity & Ambiguity

Syntactic, lexical or semantic divergence between the question and the context.

Question: Which **governing bodies** have veto power?

Context: The European Parliament and the Council of the European Union have powers of amendment and veto during the legislative process.

Multi-hop reasoning

The answer might spread across different sentences, different documents, and different logical steps.

Question: Who is Florence for Betty ?

Context: Natasha is a granddaughter to **Betty**. **Florence** is Gregorio 's sister. Gregorio is a brother of Natasha.

Missing or outdated information

The information present in the context might be outdated. The relativity and temporality of the question pose additional challenges in the current models.

Question: Who is the **current** president of Switzerland?

Context: Federal elections were held in Switzerland on 20 October **2019** to elect all members of both houses of the Federal Assembly. [...]

Recap

- Question answering is a flexible task setup used by humans in many interactions
- Question Answering can be **Open or Closed book** depending on the presence of context in the input.
- **Both generative & extractive models can be used to build QA systems.**
 - The use case of the solution (application) defines the chosen architecture.
- **Evaluation of the output** depends on the task and can be very challenging.

References

- Rajpurkar, Pranav, et al. "Squad: 100,000+ questions for machine comprehension of text." *arXiv preprint arXiv:1606.05250* (2016).
- Kwiatkowski, Tom, et al. "Natural questions: a benchmark for question answering research." *Transactions of the Association for Computational Linguistics* 7 (2019): 453-466.

Retrieval-Augmented LMs

Antoine Bosselut



Reading Comprehension challenges

Factuality

Temporality

AN 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.

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

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

FEVER 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

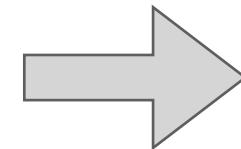
Today's Outline

- **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

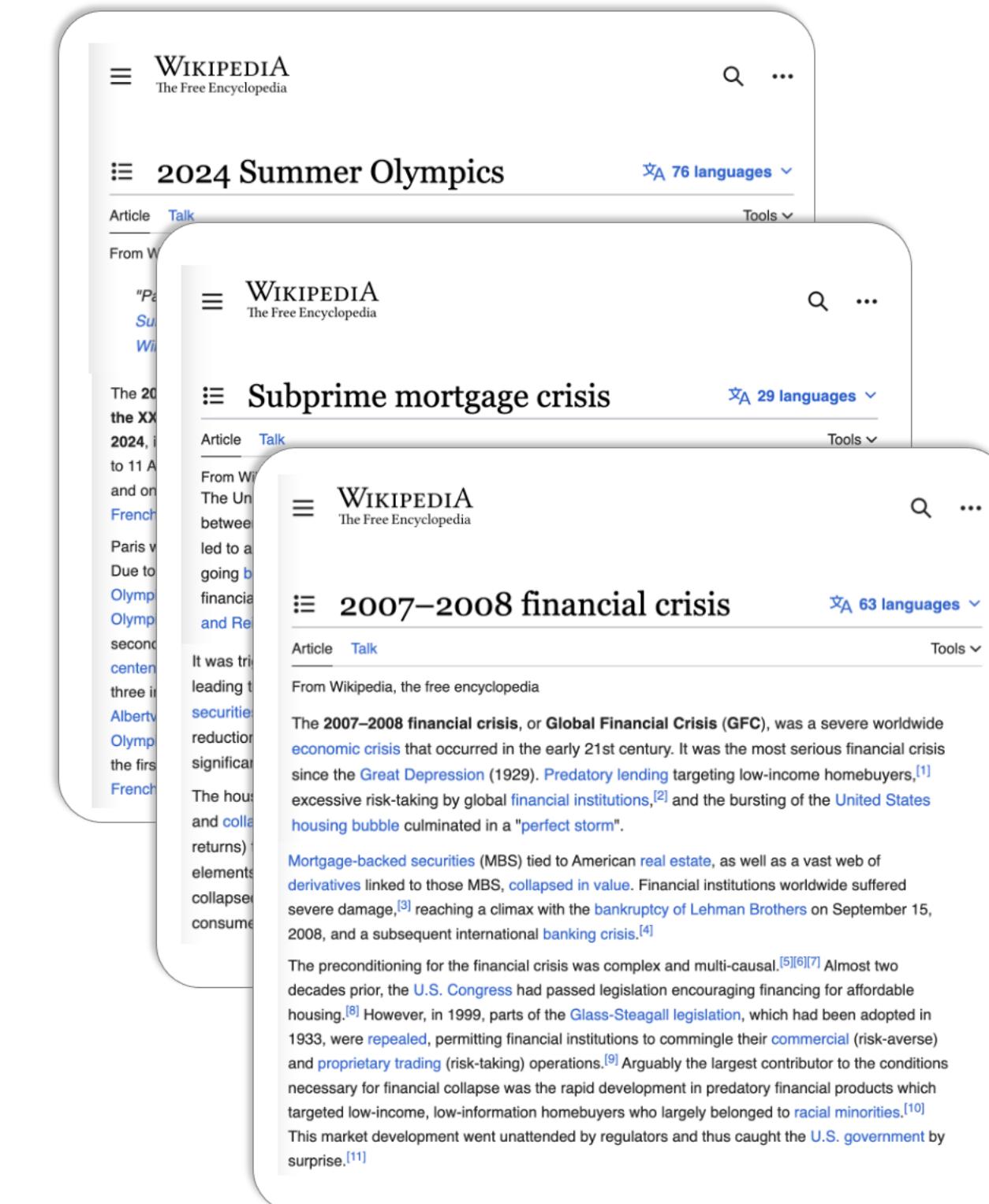
Finding the answer in 21M documents

Query



Documents

"Where the financial crisis of 2008 started?"



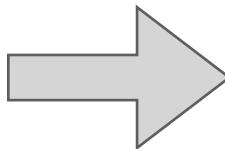
**Why can't we do this with closed-book
Extractive & Generative QA models?**

We don't have a context!

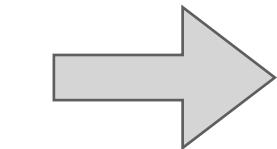
But we can get one!
(With adaptations)

Finding the answer in 21M documents

Query



Documents



Retrieve relevant documents

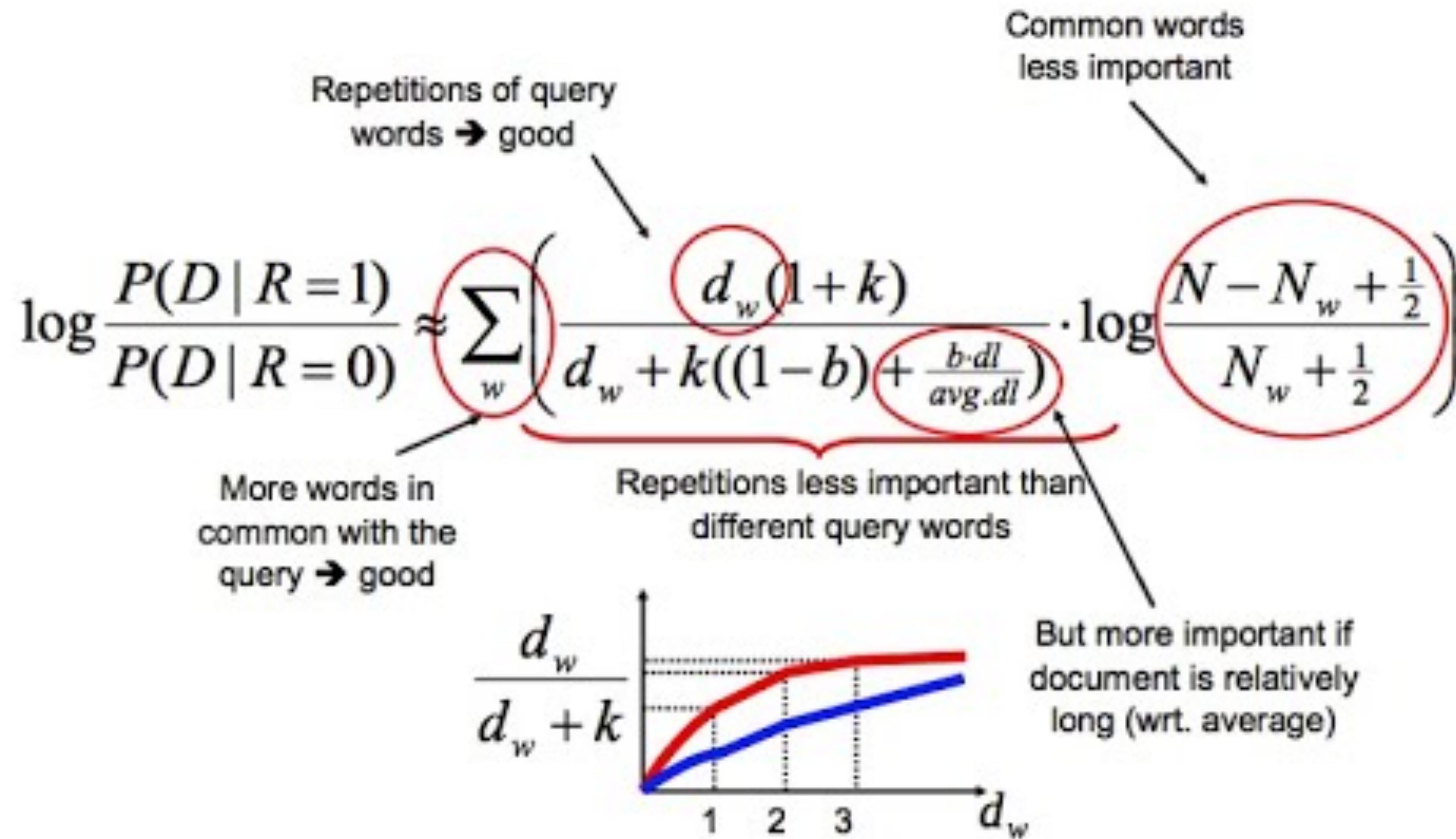
That might contain the answer

"Where the financial crisis of 2008 started?"

Three overlapping screenshots of a Wikipedia mobile interface. The top screenshot shows the search results for "2024 Summer Olympics". The middle screenshot shows the search results for "Subprime mortgage crisis". The bottom screenshot shows the search results for "2007–2008 financial crisis". Each screenshot displays a list of articles related to the query, with the main article's title and summary visible.

A single screenshot of a Wikipedia mobile interface showing the search result for "2007–2008 financial crisis". The page title is "2007–2008 financial crisis" and it includes a summary, article text, and external links.

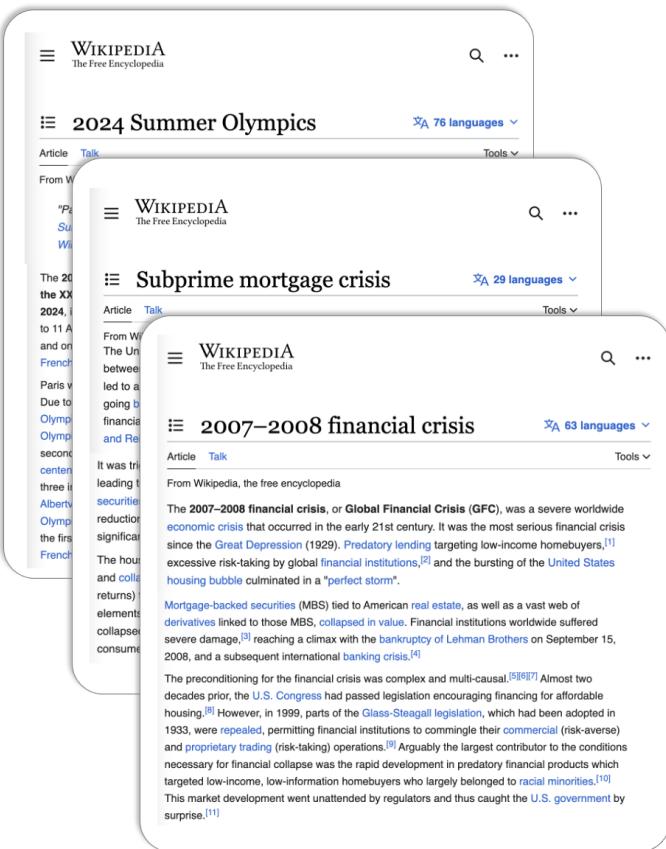
Classical Retrieval: Okapi BM25



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

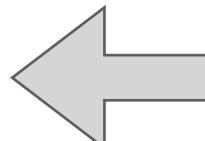
Dense Passage Retrieval (DPR)

Documents



Query

Dense Embedding Model

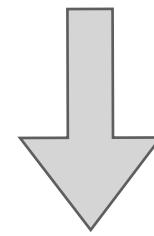


"Where the financial crisis of 2008 started?"

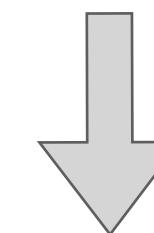
- Create the representations of documents
- Create the representation of the query
- Retrieve k documents vectors based on the query vector

Dense Passage Retrieval (DPR)

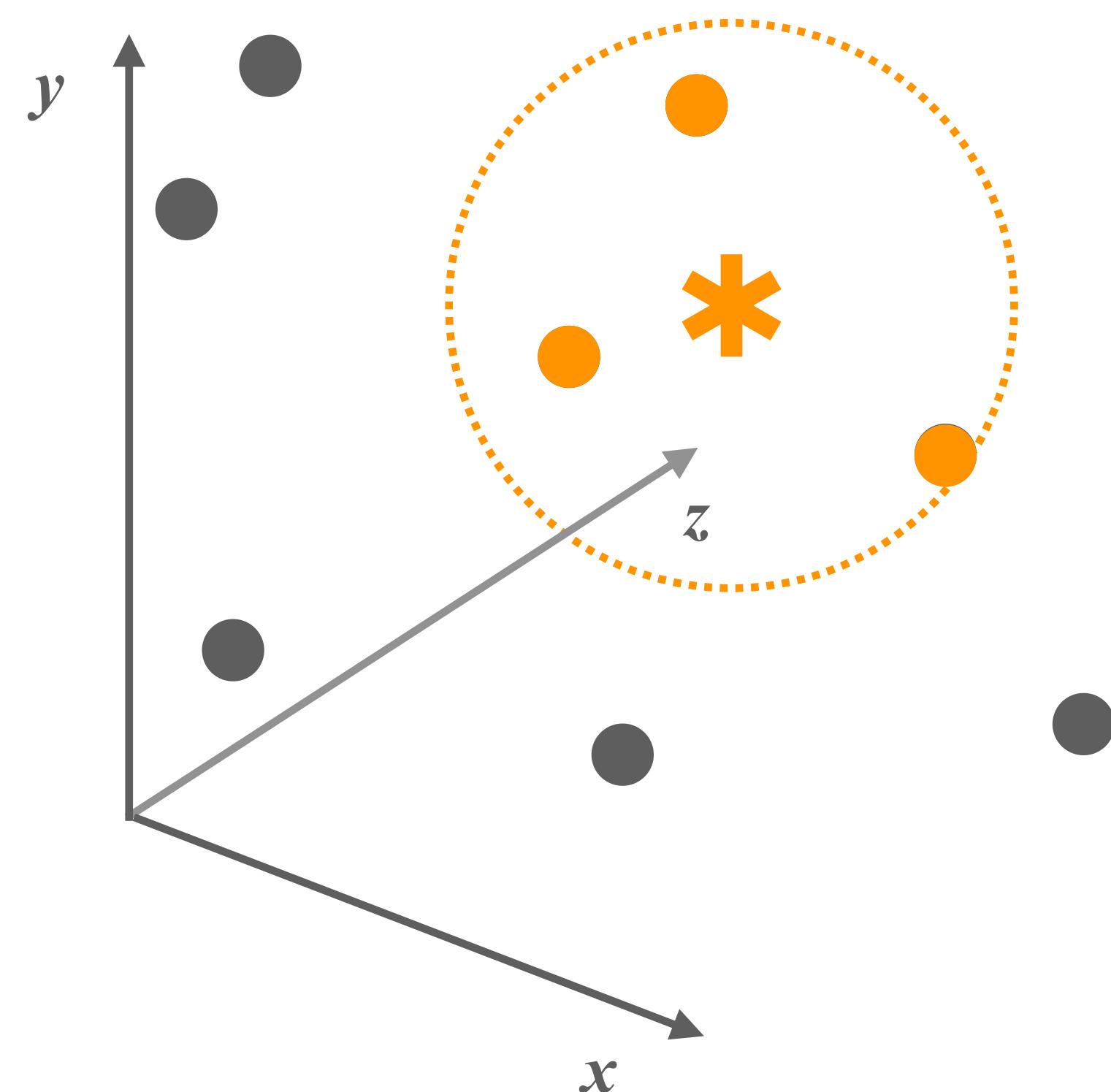
Documents



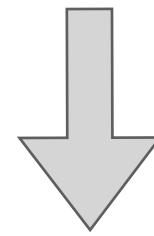
$E_P(\cdot)$



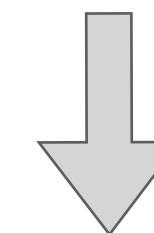
`[-0.5968882, -0.33086956, -0.32643065, -0.3670732, ...]`



Query



$E_Q(\cdot)$



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

$$\text{sim}(q, p) = E_Q(q)^\top E_P(p)$$

Training DPR

"Where the financial crisis of 2008 started?"

How to create a Document-Query vector space?

Goal: Relevant pairs of questions-passages will have a smaller distance compared to the irrelevant ones.

Positive passage p^+



Negative passage p^-



DPR LOSS FUNCTION

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

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

Retrieval-Augmented LMs

$$p(y | x) =$$

LM

Retriever

Auto-Encoder

$$\sum_{z \in \mathcal{Z}}$$

$$p(y | z, x)$$

Auto-Regressive

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

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

$$p(z | x)$$

Auto-Encoder

Trained to retrieve relevant
documents (optional)

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

Retrieval-Augmented LMs - Terminology

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

memory

knowledge

The type of external source the retriever will use.

modalities

Implicit

Parametric

LM

vs

Explicit

Non-parametric

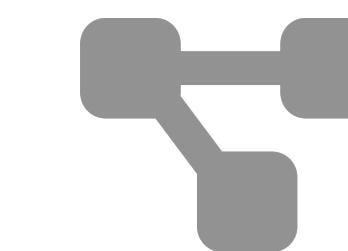
vs

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

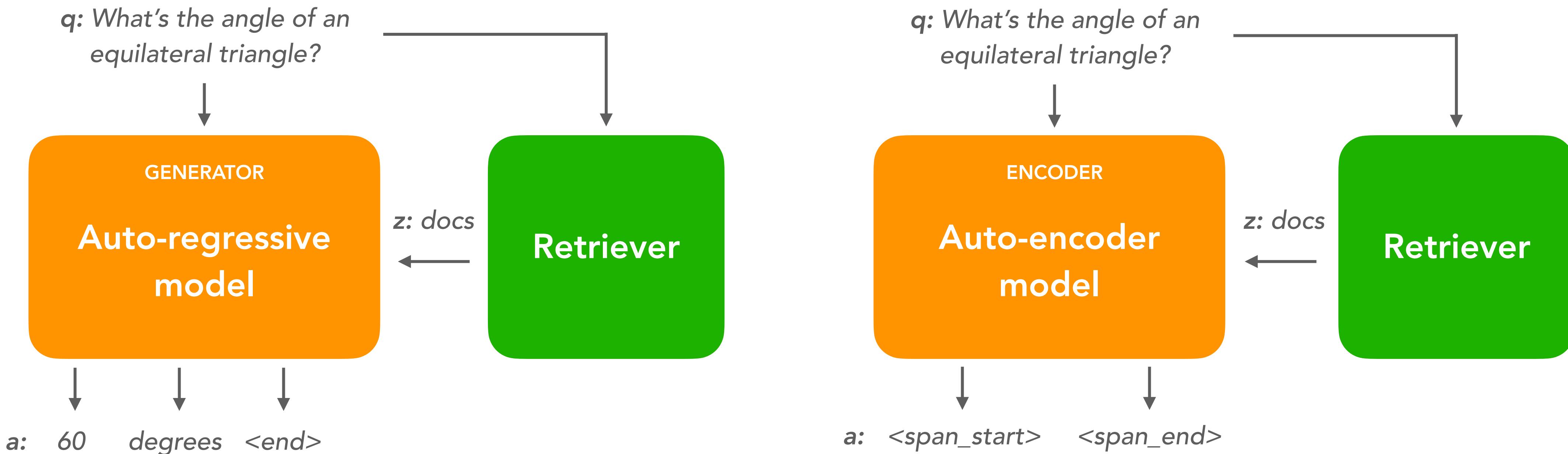
REALM: Pre-training & KB

TYPES OF EXTERNAL KNOWLEDGE

Document Repositories
Knowledge Graphs

ERNIE: Pre-training & KG

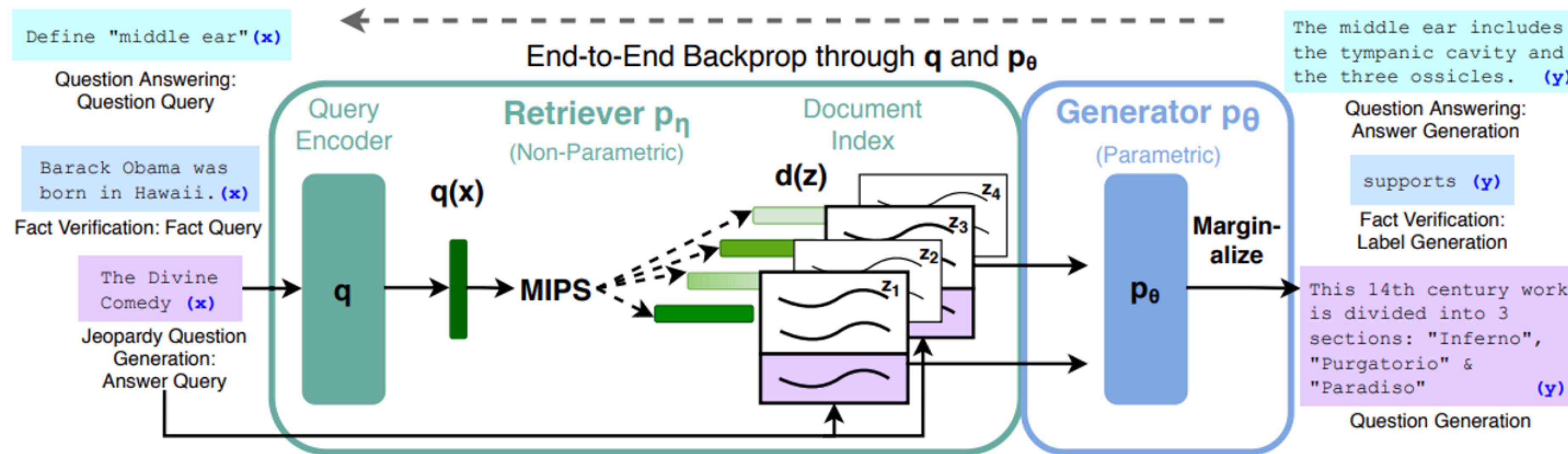
Generative vs Extractive



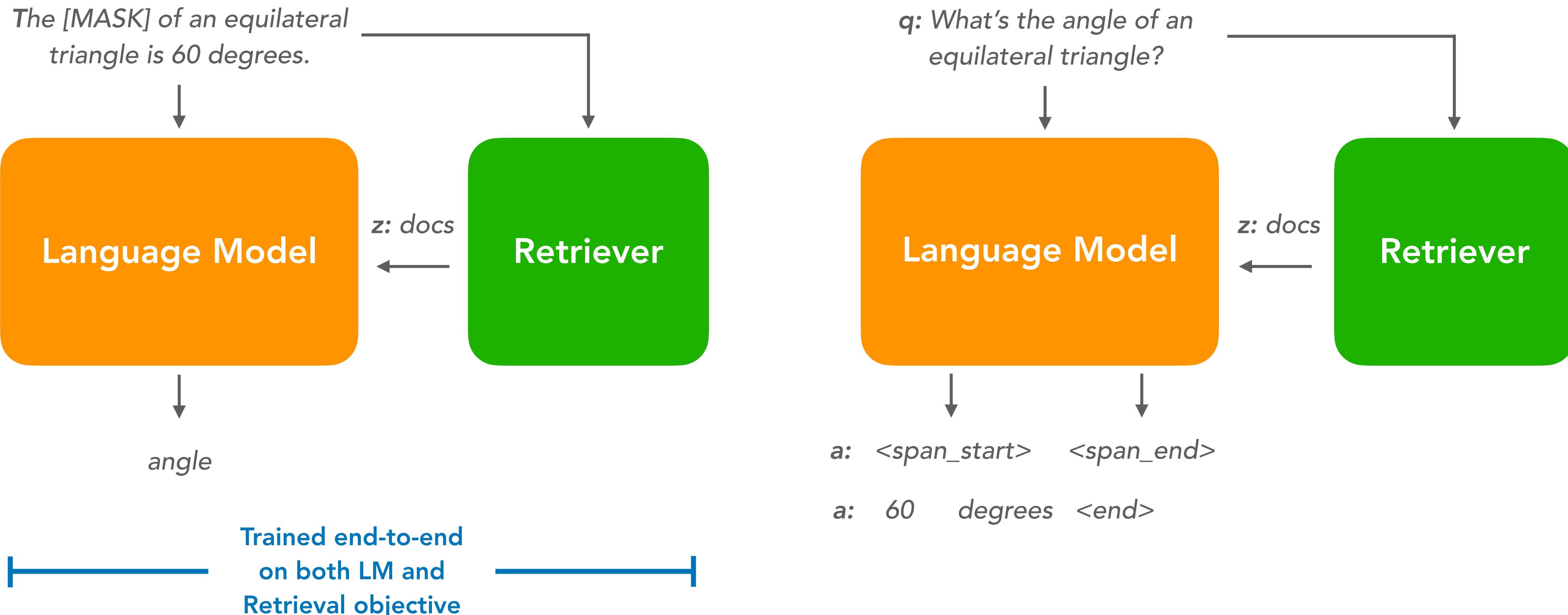
RAG: Generative Retrieval-Augmented LM

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

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



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

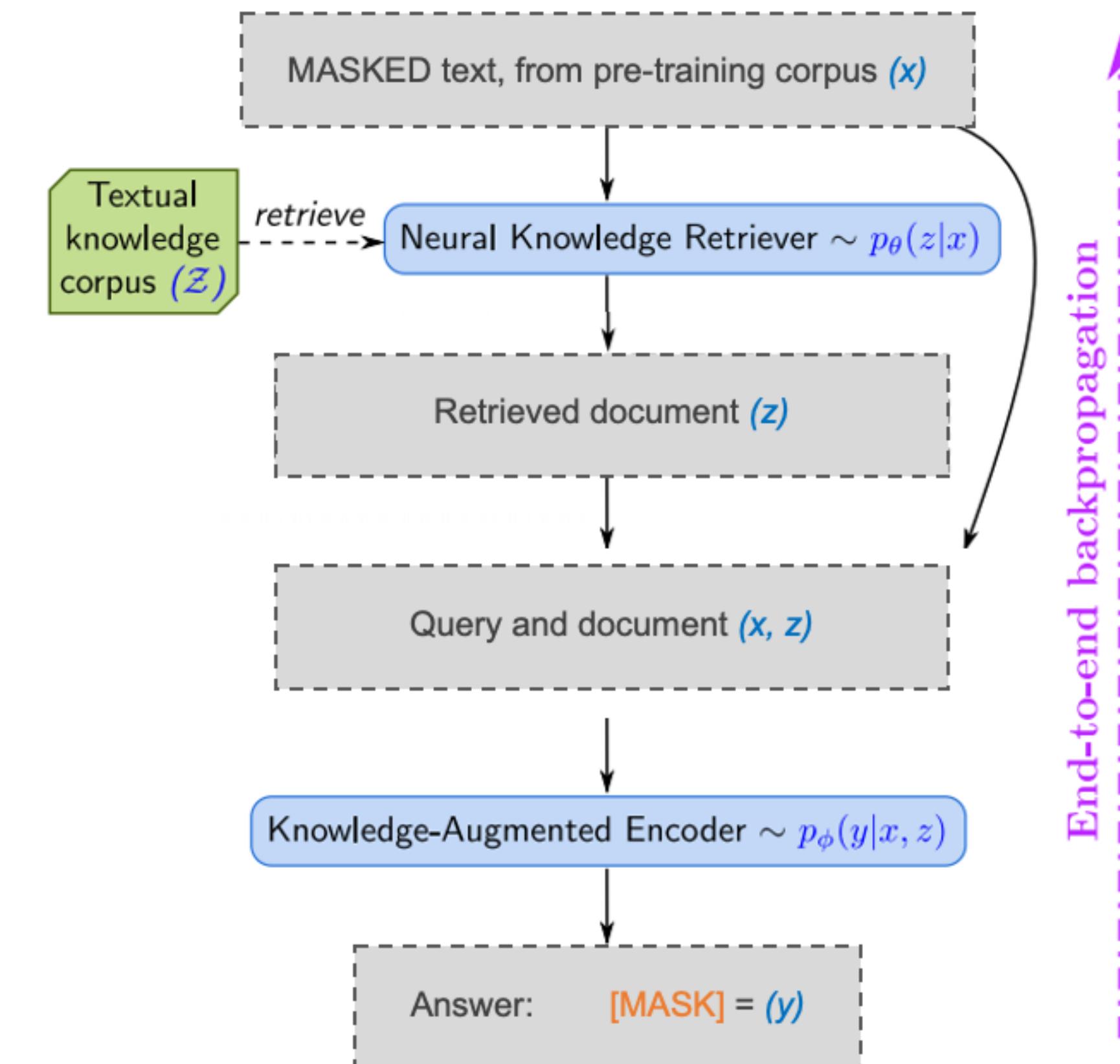
Then Predict:

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

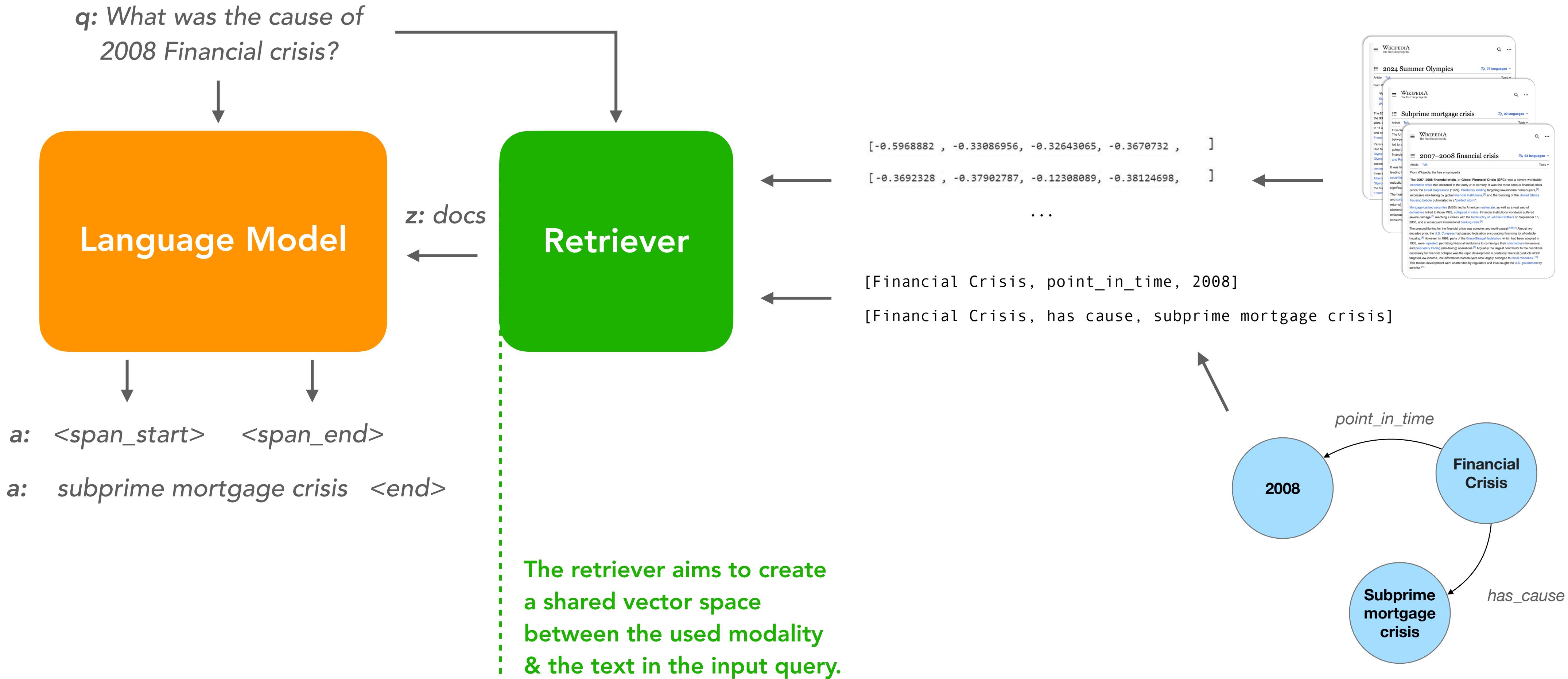
Goal: Minimise perplexity

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.



Different types of external knowledge

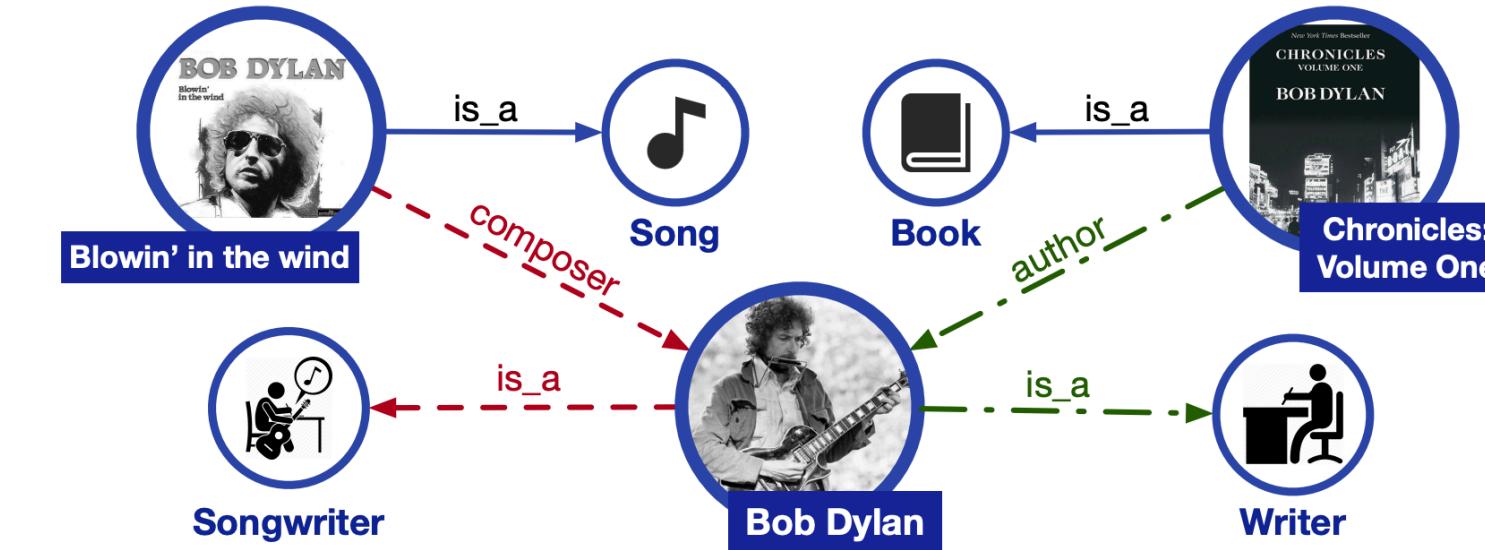


What other modalities could we use as a base ?

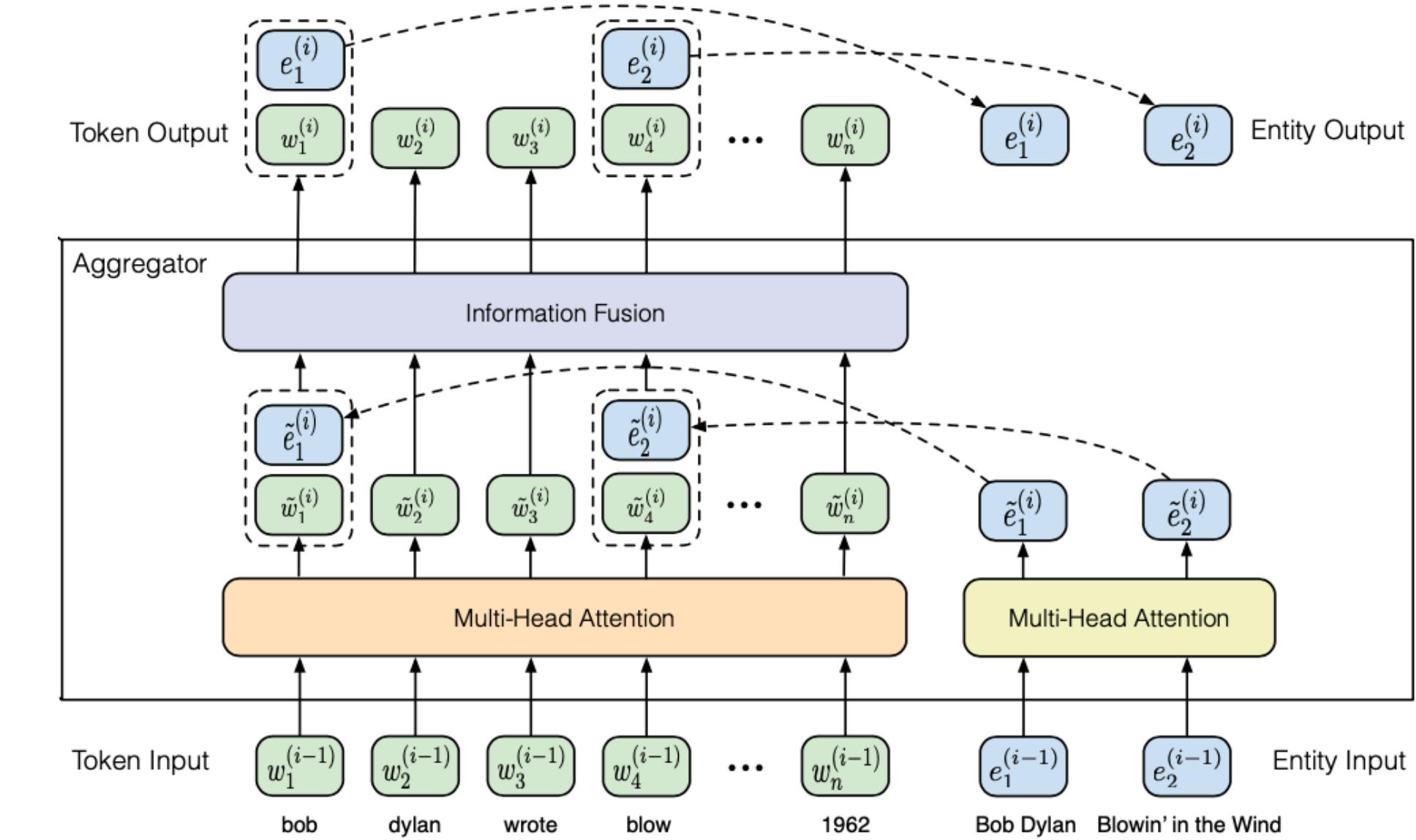
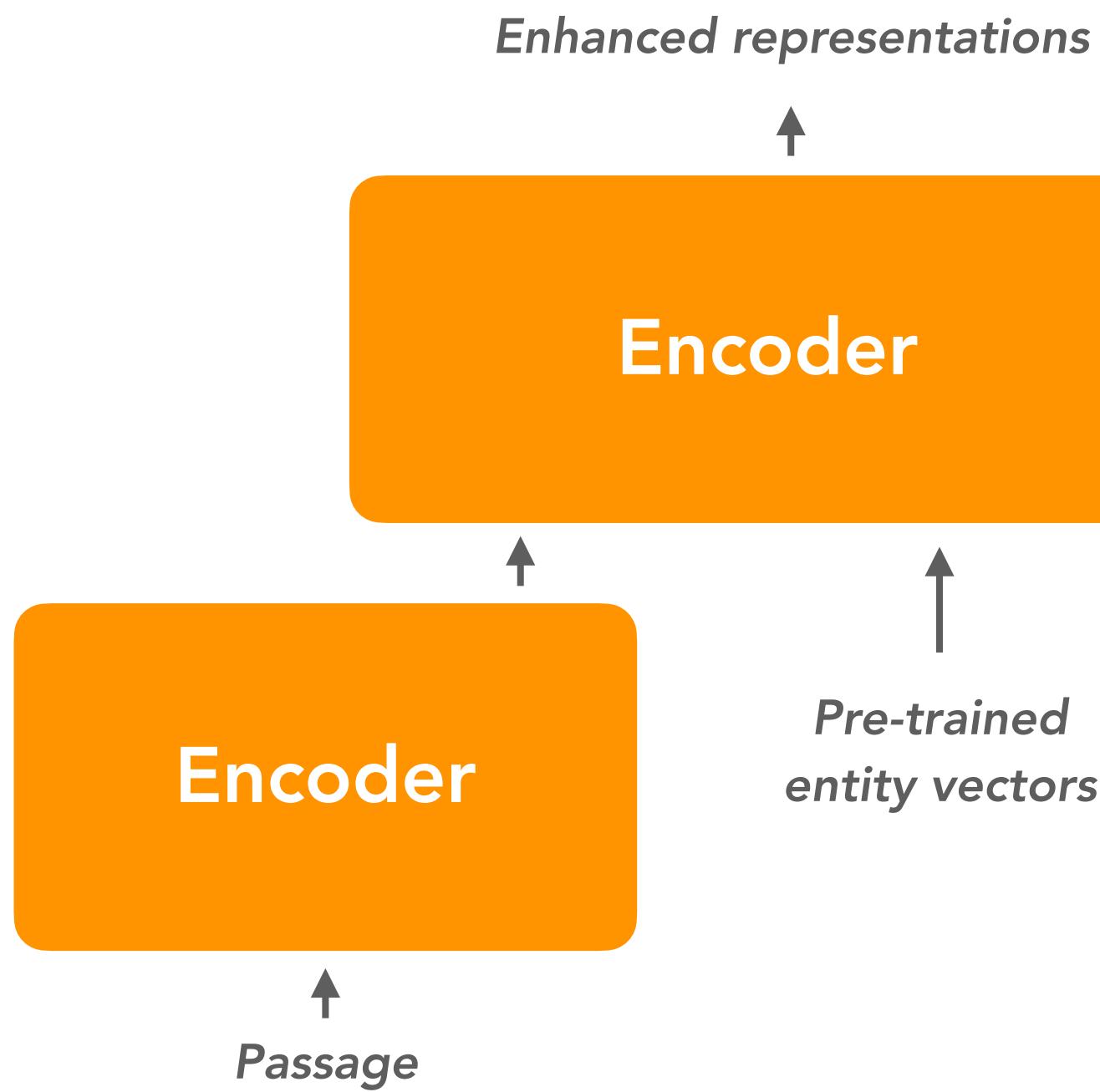
How would we integrate these modalities ?

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

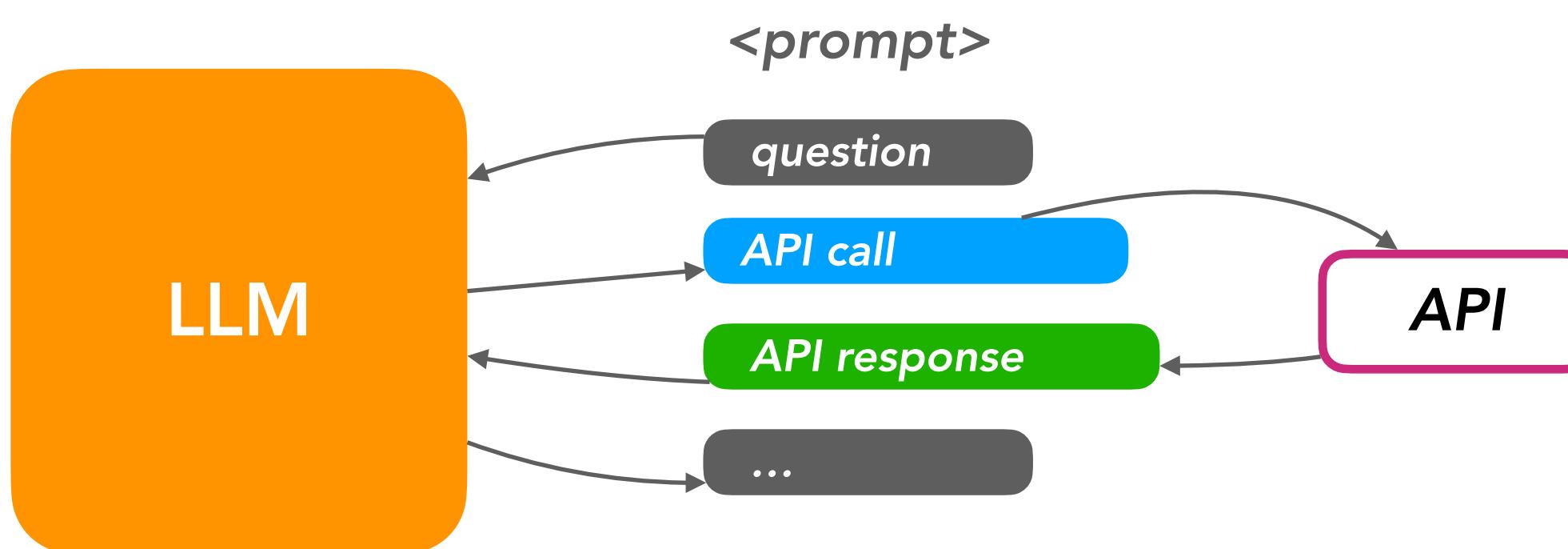
We looked at retrieving information from document bases and knowledge graphs.

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.

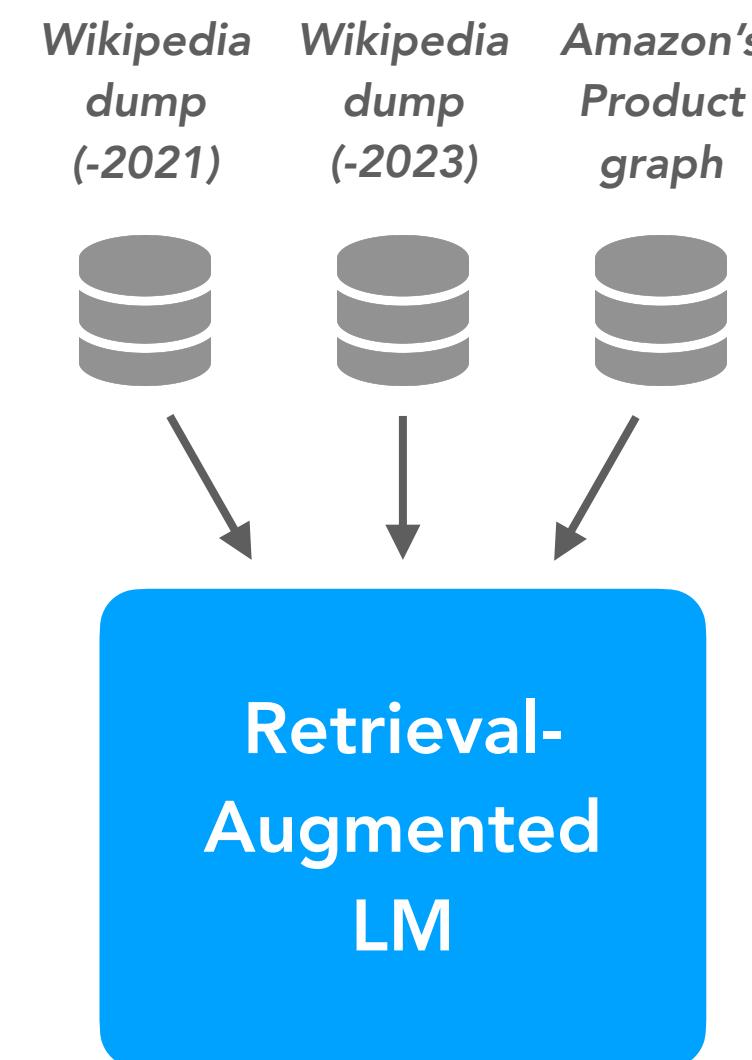
Toolformer (Schick et al. 2023)

What are some benefits of augmentation?

Additional benefits of Augmented LMs

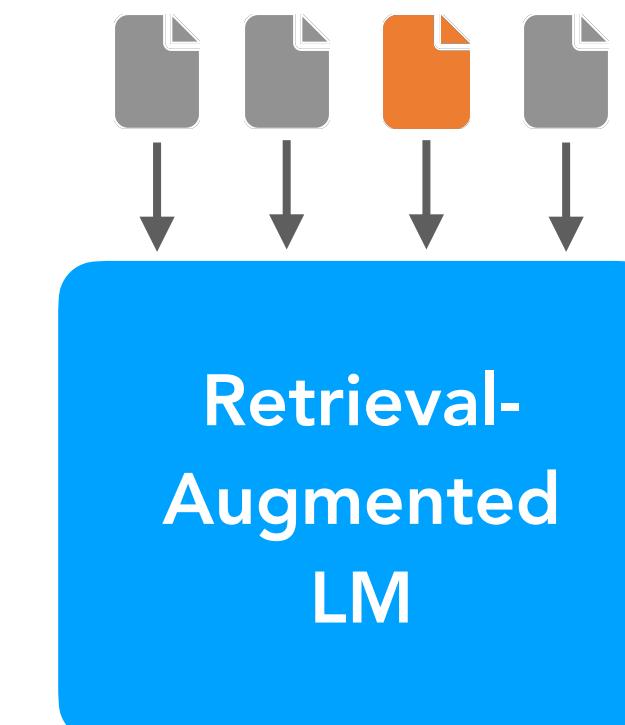
Modularity

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



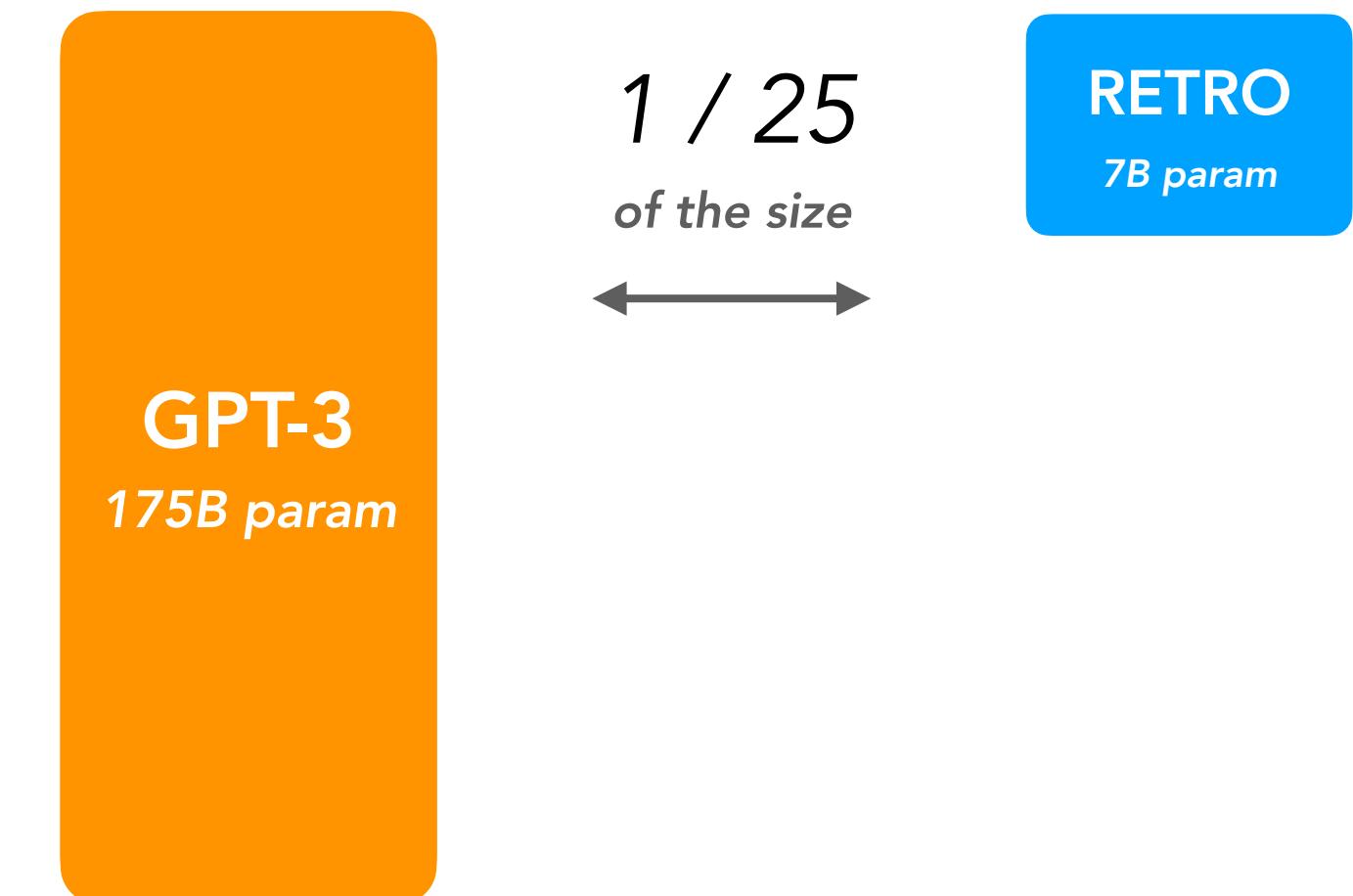
Attribution

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	
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	
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).