# Question Answering & Reading Comprehension

Antoine Bosselut

EPFL

nlp lab

# Announcements

- **No Lecture Tomorrow!**
- **Course Project:** Milestone 2 due **Sunday, May 28th**!
  - Data from interactions released today!
- **Internship opportunity:** MLO and NLP labs hosting multiple interns to work on training LLMs this summer:
  - Send CV and transcript to: **nlp-mlo-llm-internship-apply@groupes.epfl.ch**
  - Interviews already starting now. **Send application ASAP if interested.**

# Today's Outline

- **Lecture**
  - **Question Answering:** Tasks, Models, Limitations
- **Guest Lecture:** Angelika Romanou
  - **Infuse 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**

## Where do we find QA systems?

# 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:

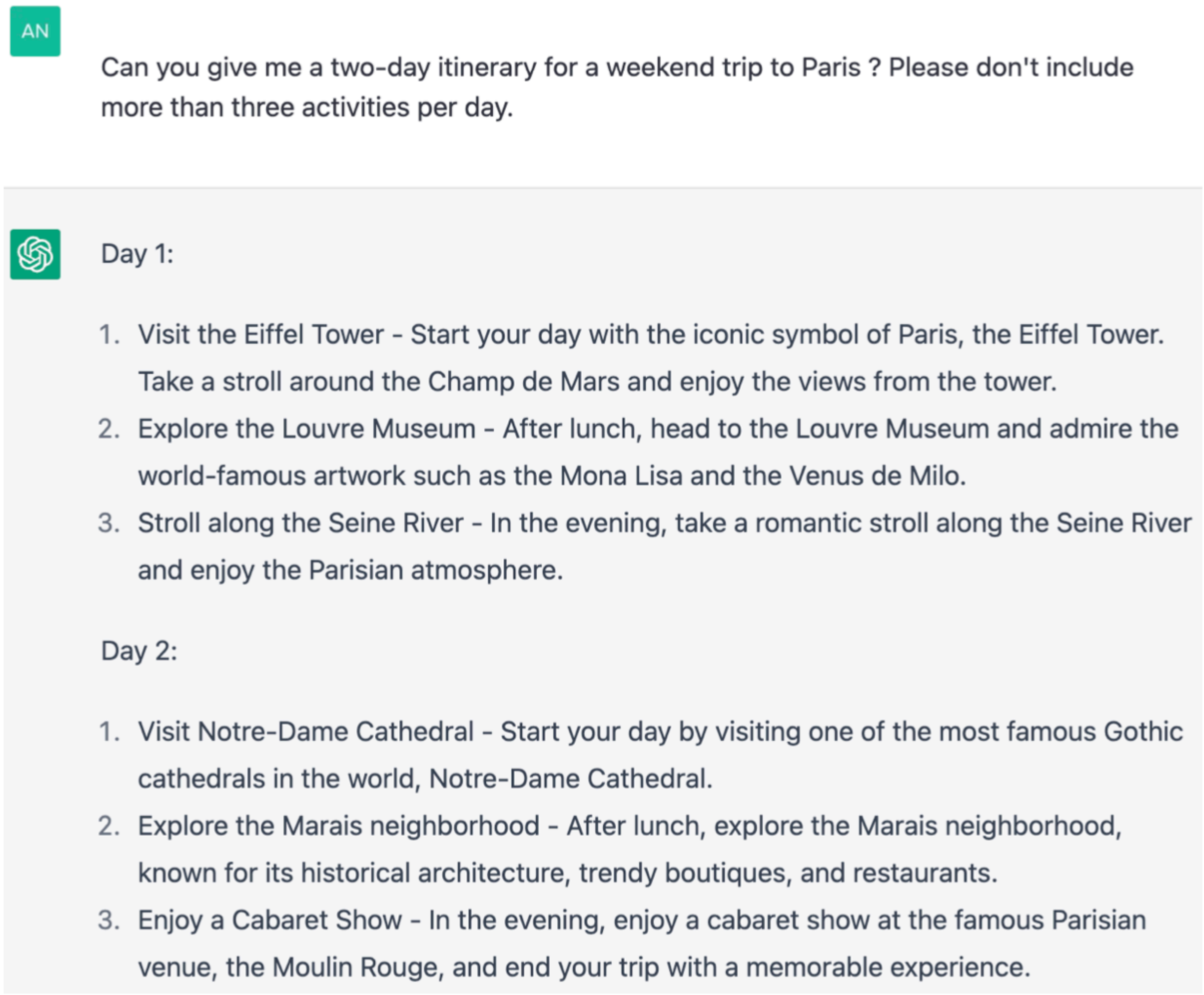
1. Visit the Eiffel Tower - Start your day with the iconic symbol of Paris, the Eiffel Tower. Take a stroll around the Champ de Mars and enjoy the views from the tower.
2. Explore the Louvre Museum - After lunch, head to the Louvre Museum and admire the world-famous artwork such as the Mona Lisa and 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.

# Question Answering Systems

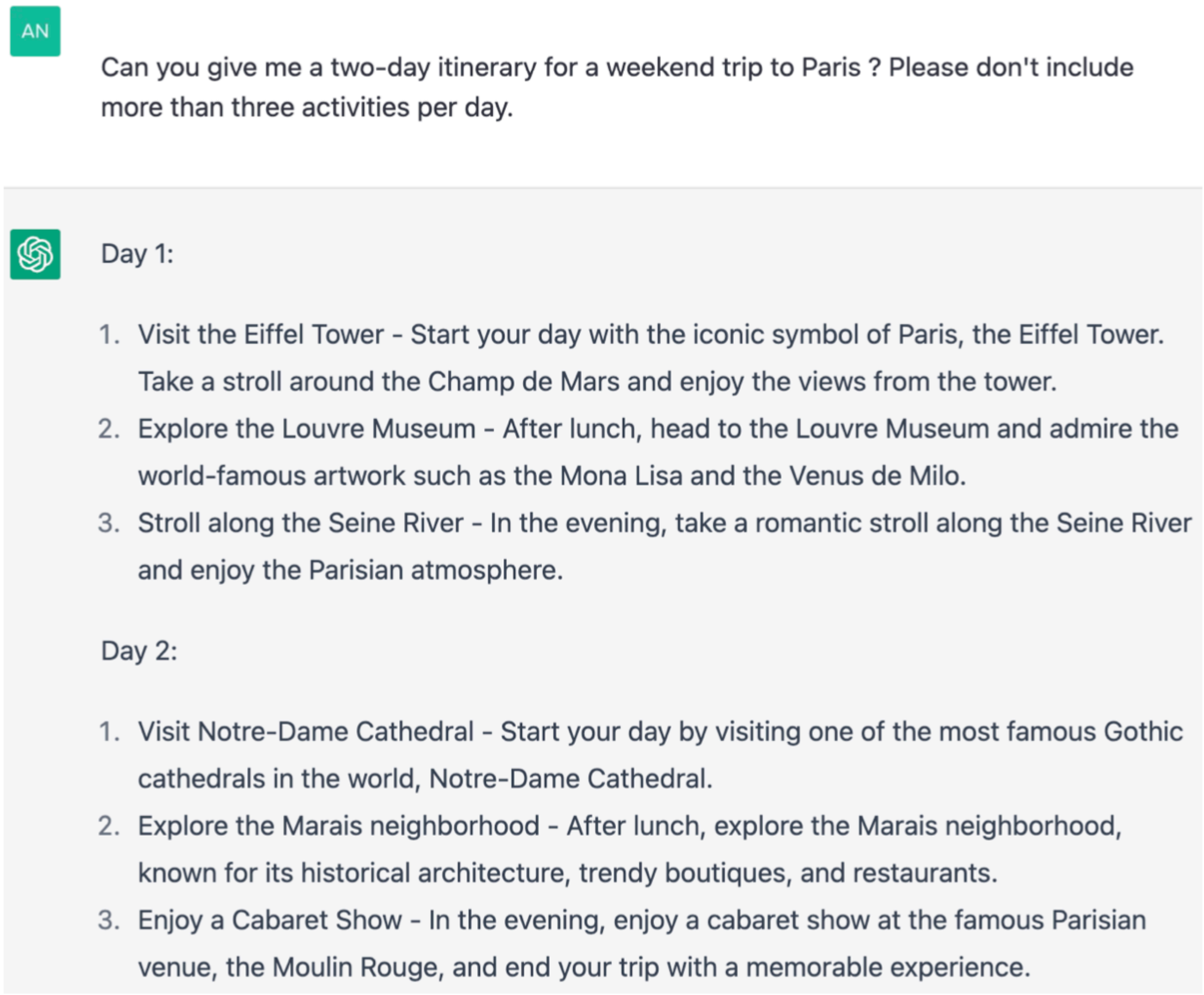
## Chatbots




## Search Engines

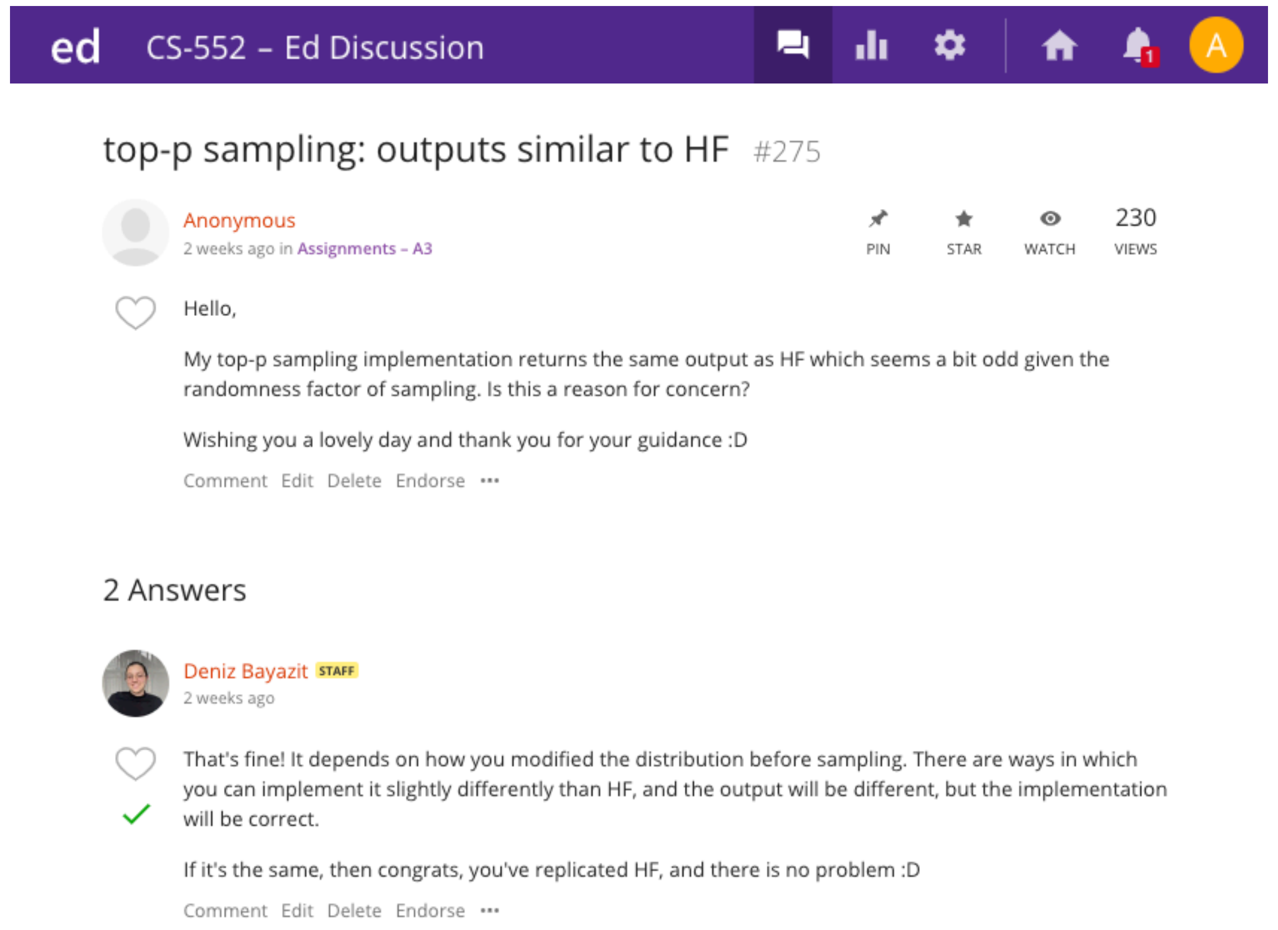
# Question Answering Systems

## Chatbots




## Search Engines


## Forums / e-learning portals



# 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 English?

# 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 English?

**Output**

## Answer

Amazonia

# 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 English?

**Question Answering Model**

**Output**

### Answer

Amazonia

How might we have historically designed a QA system?

# Classical QA

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

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

$$
\begin{array}{cc}
\overline{\quad\quad\quad\quad\quad\quad\quad\quad}^{>} & \overline{\quad\quad\quad\quad\quad\quad\quad\quad}^{>} \\
N & N\backslash N \\
\lambda x. former(x) \land municipalities(x) & \lambda f \lambda y. f(y) \land in(y, Brandenburg)
\end{array}
$$

$$
\begin{array}{c}
\overline{\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad}^{<} \\
N \\
l_0 = \lambda x. former(x) \land municipalities(x) \land in(x, Brandenburg)
\end{array}
$$

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

$$\mathbf{l}_0 = \lambda x. former(x) \land municipalities(x) \land in(x, Brandenburg)$$
$$\mathbf{l}_1 = \lambda x. former(x) \land municipalities(x) \land in(x, \texttt{Brandenburg})$$
$$\mathbf{l}_2 = \lambda x. former(x) \land municipalities(x) \land \texttt{location.containedby}(x, \texttt{Brandenburg})$$
$$\mathbf{l}_3 = \lambda x. former(x) \land \texttt{OpenRel}(x, \texttt{Municipality}) \land \texttt{location.containedby}(x, \texttt{Brandenburg})$$
$$\mathbf{l}_4 = \lambda x. \texttt{OpenType}(x) \land \texttt{OpenRel}(x, \texttt{Municipality}) \land \texttt{location.containedby}(x, \texttt{Brandenburg})$$

- 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.

| Former | municipalities | in | Brandenburgh |
|---|---|---|---|
| $N/N$ | $N$ | $N\backslash N/NP$ | $NP$ |
| $\lambda f \lambda x.f(x) \wedge former(x)$ | $\lambda x.municipalities(x)$ | $\lambda f \lambda x \lambda y.f(y) \wedge in(y, x)$ | $Brandenburg$ |

$$\frac{\lambda f \lambda x.f(x) \wedge former(x) \quad \lambda x.municipalities(x)}{N : \lambda x.former(x) \wedge municipalities(x)}>$$

$$\frac{\lambda f \lambda x \lambda y.f(y) \wedge in(y, x) \quad Brandenburg}{N\backslash N : \lambda f \lambda y.f(y) \wedge in(y, Brandenburg)}>$$

$$\frac{\lambda x.former(x) \wedge municipalities(x) \quad \lambda f \lambda y.f(y) \wedge in(y, Brandenburg)}{N : l_0 = \lambda x.former(x) \wedge municipalities(x) \wedge in(x, Brandenburg)}<$$

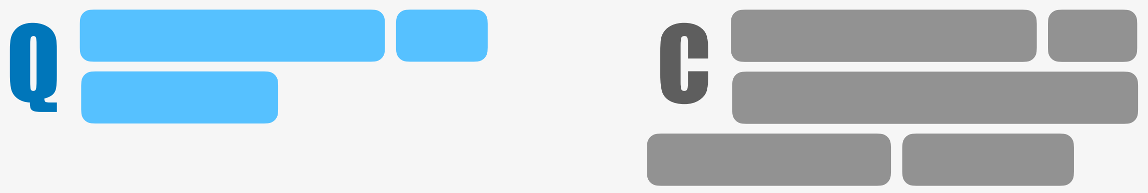
(b) **Constant matches** replace underspecified constants with Freebase concepts

$$\mathbf{l_0} = \lambda x.former(x) \wedge municipalities(x) \wedge in(x, Brandenburg)$$
$$\mathbf{l_1} = \lambda x.former(x) \wedge municipalities(x) \wedge in(x, \texttt{Brandenburg})$$
$$\mathbf{l_2} = \lambda x.former(x) \wedge municipalities(x) \wedge \texttt{location.containedby}(x, \texttt{Brandenburg})$$
$$\mathbf{l_3} = \lambda x.former(x) \wedge \texttt{OpenRel}(x, \texttt{Municipality}) \wedge \texttt{location.containedby}(x, \texttt{Brandenburg})$$
$$\mathbf{l_4} = \lambda x.\texttt{OpenType}(x) \wedge \texttt{OpenRel}(x, \texttt{Municipality}) \wedge \texttt{location.containedby}(x, \texttt{Brandenburg})$$

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

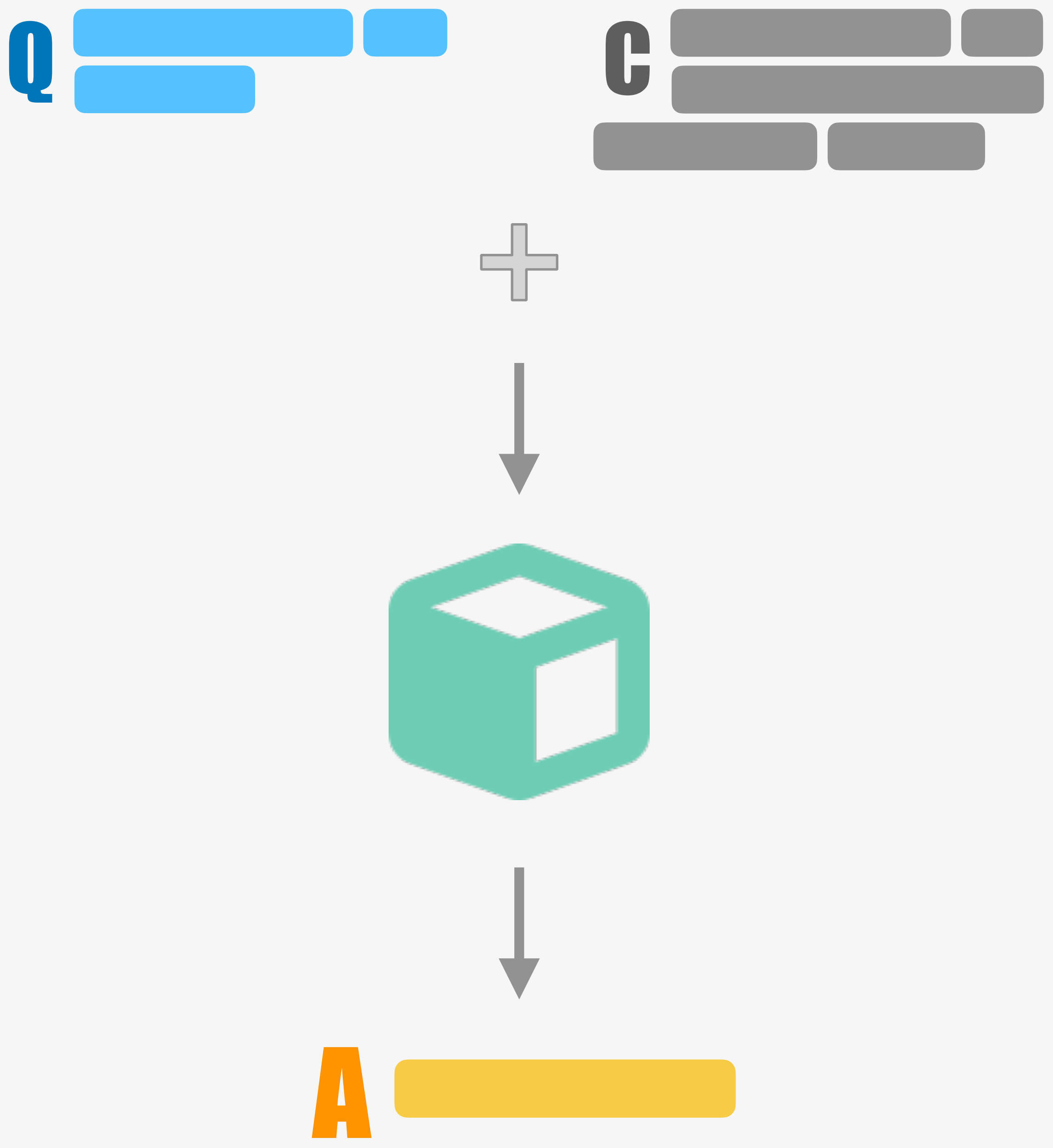
# Types of QA

## Extractive QA



# Types of QA

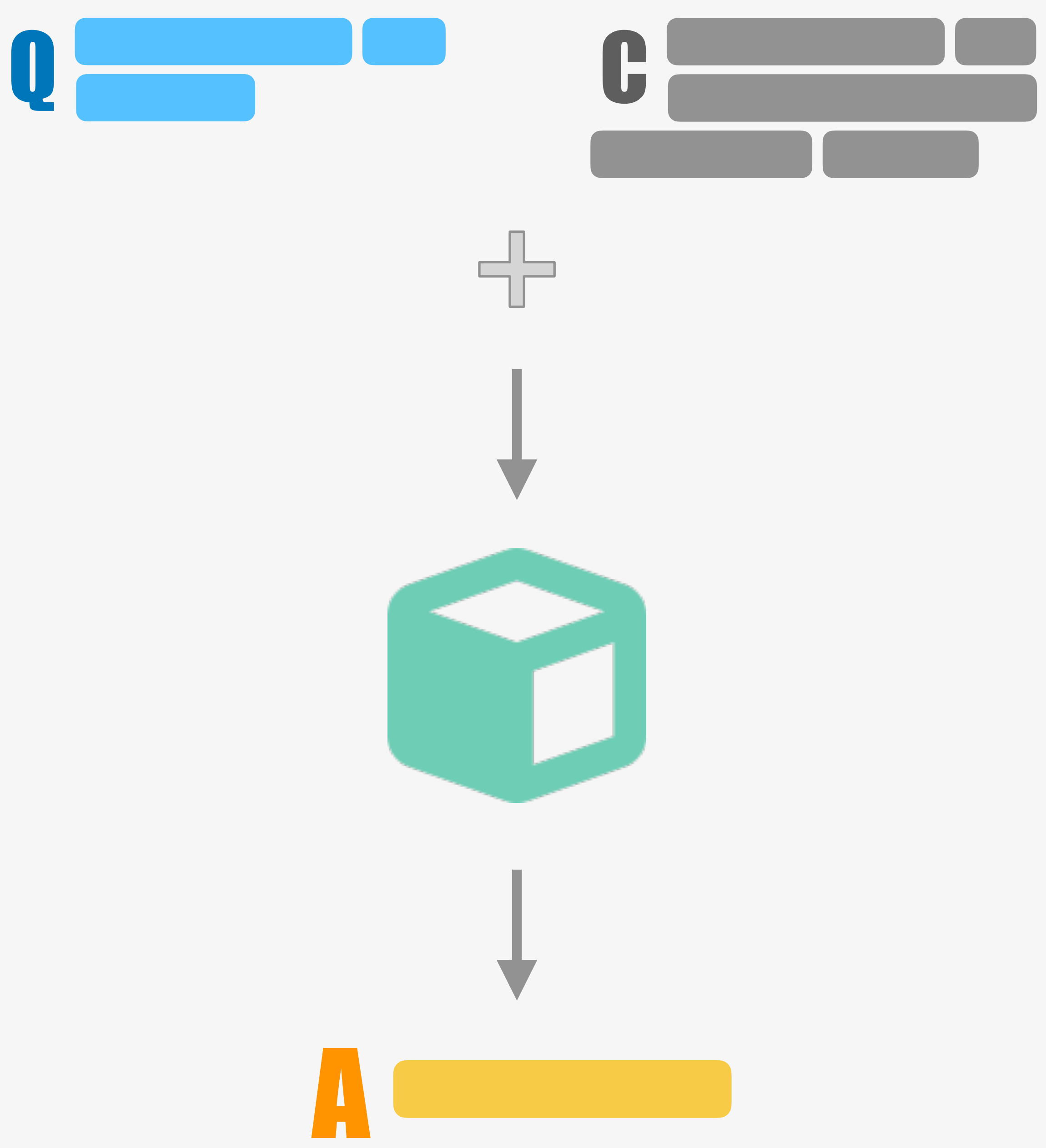
## Extractive QA




Answer is extracted
from the context

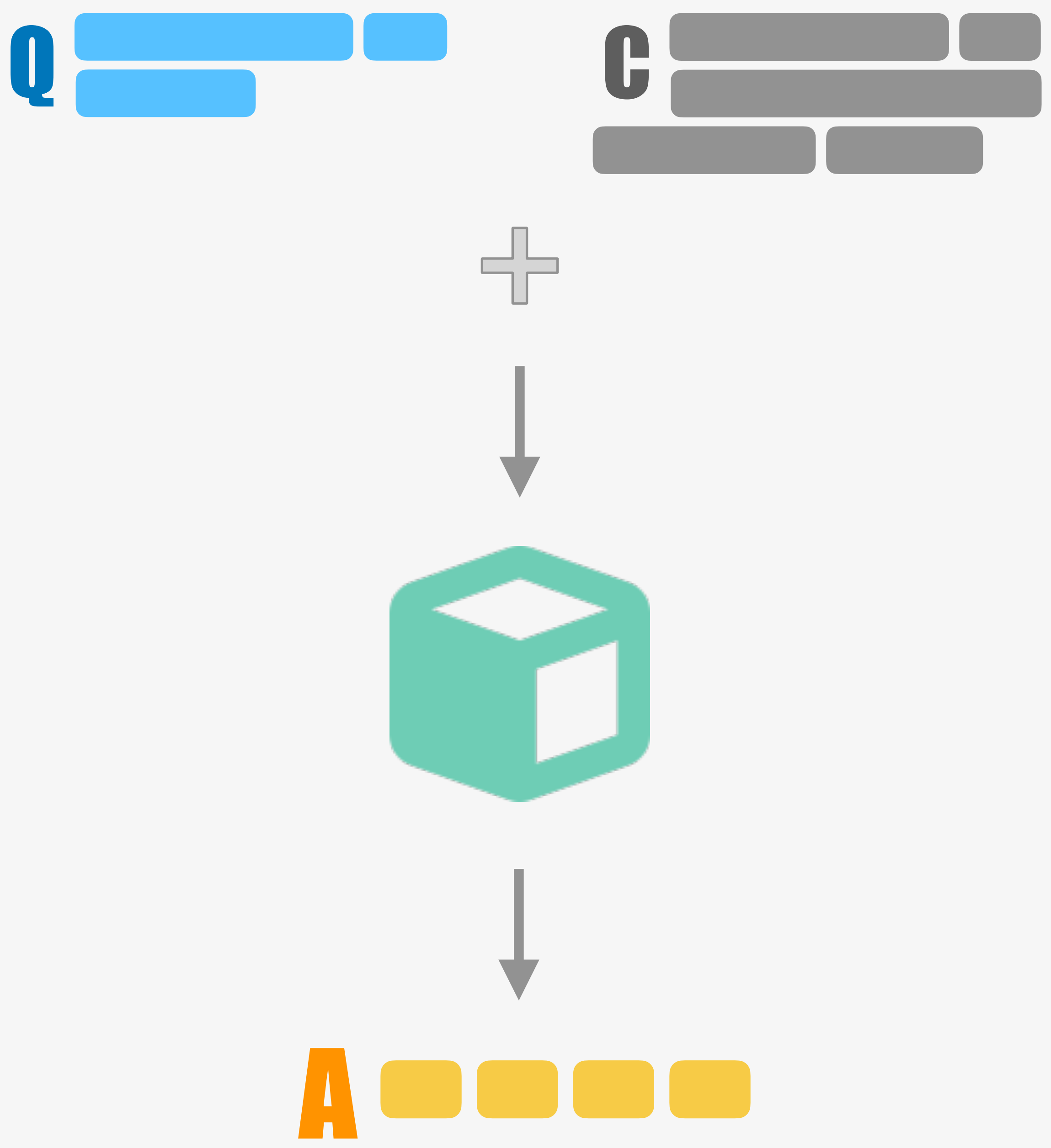
# Types of QA

## Extractive QA




Answer is extracted
from the context

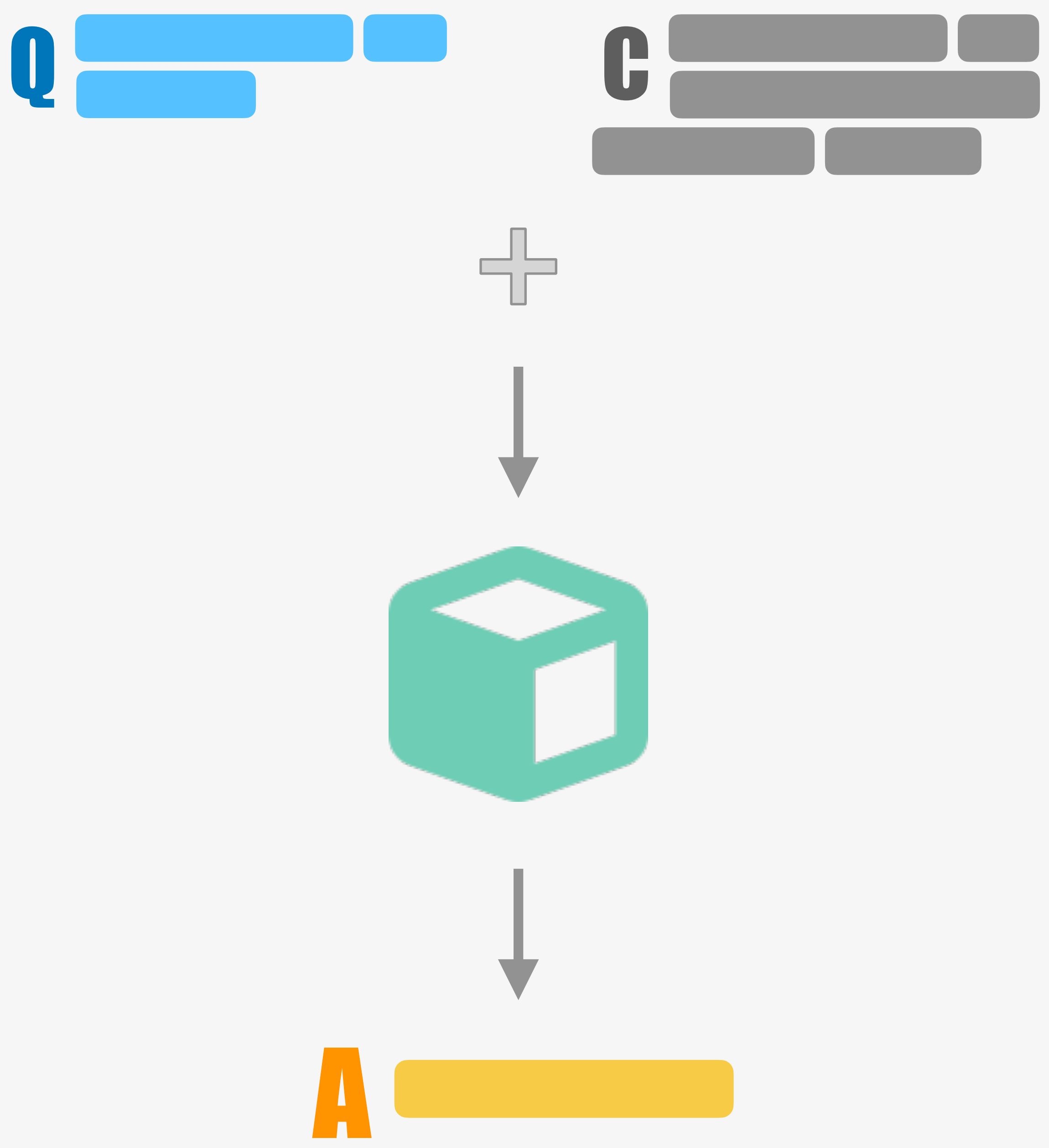
## Open-Generative QA




Answer is generated
in an auto-regressive way

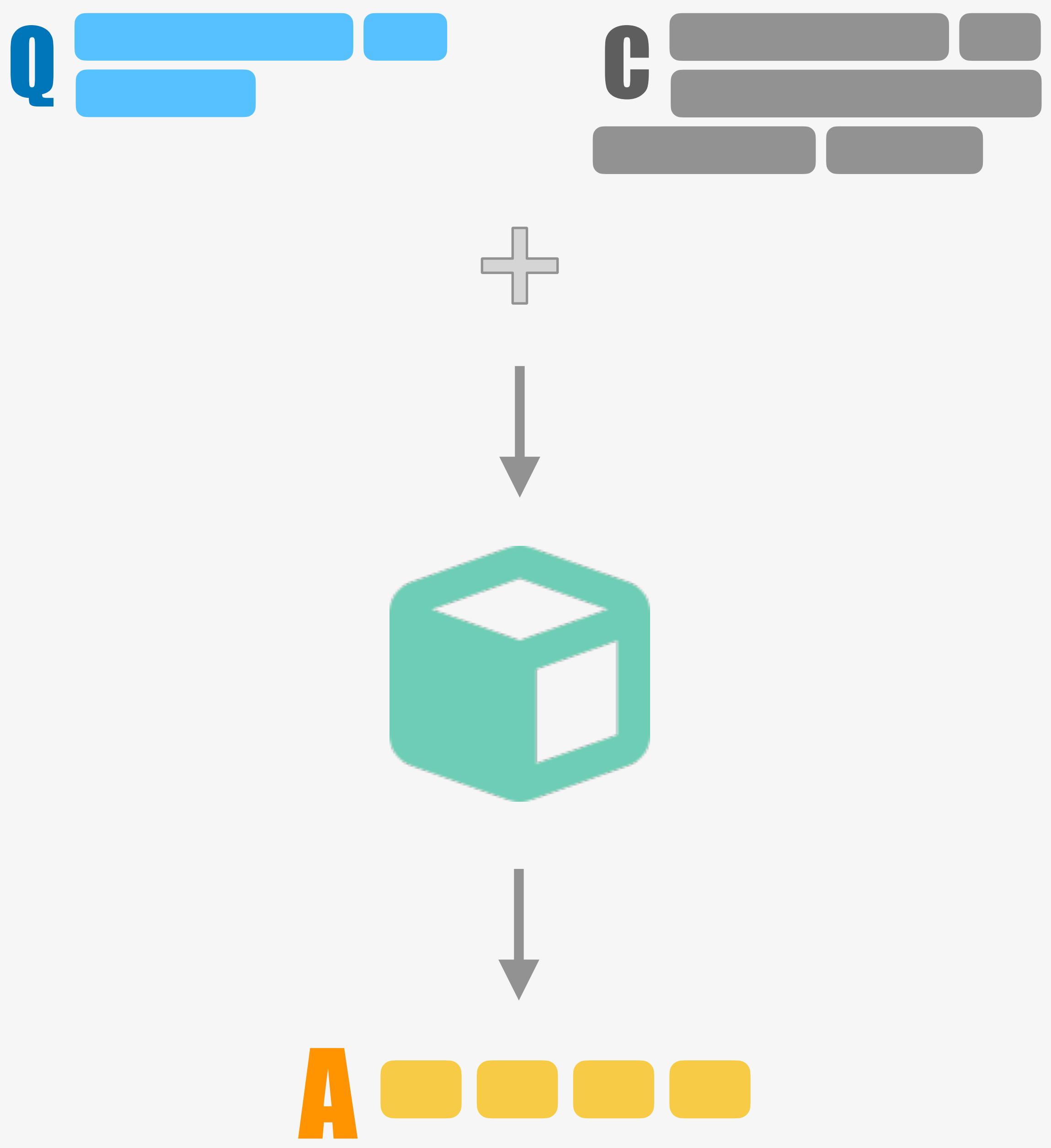
# Types of QA

## Extractive QA



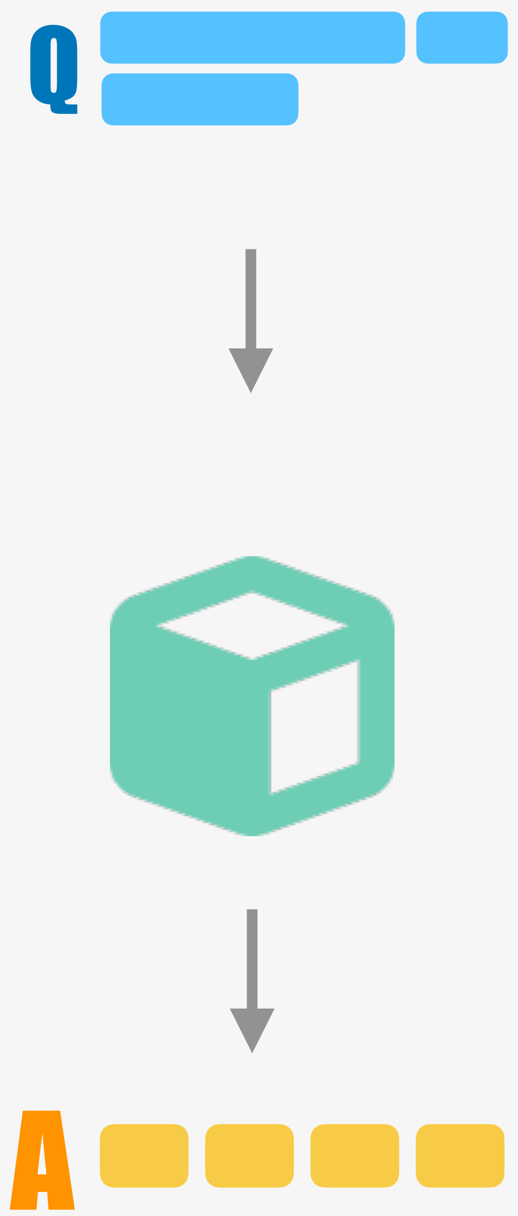

Answer is extracted
from the context

## Open-Generative QA



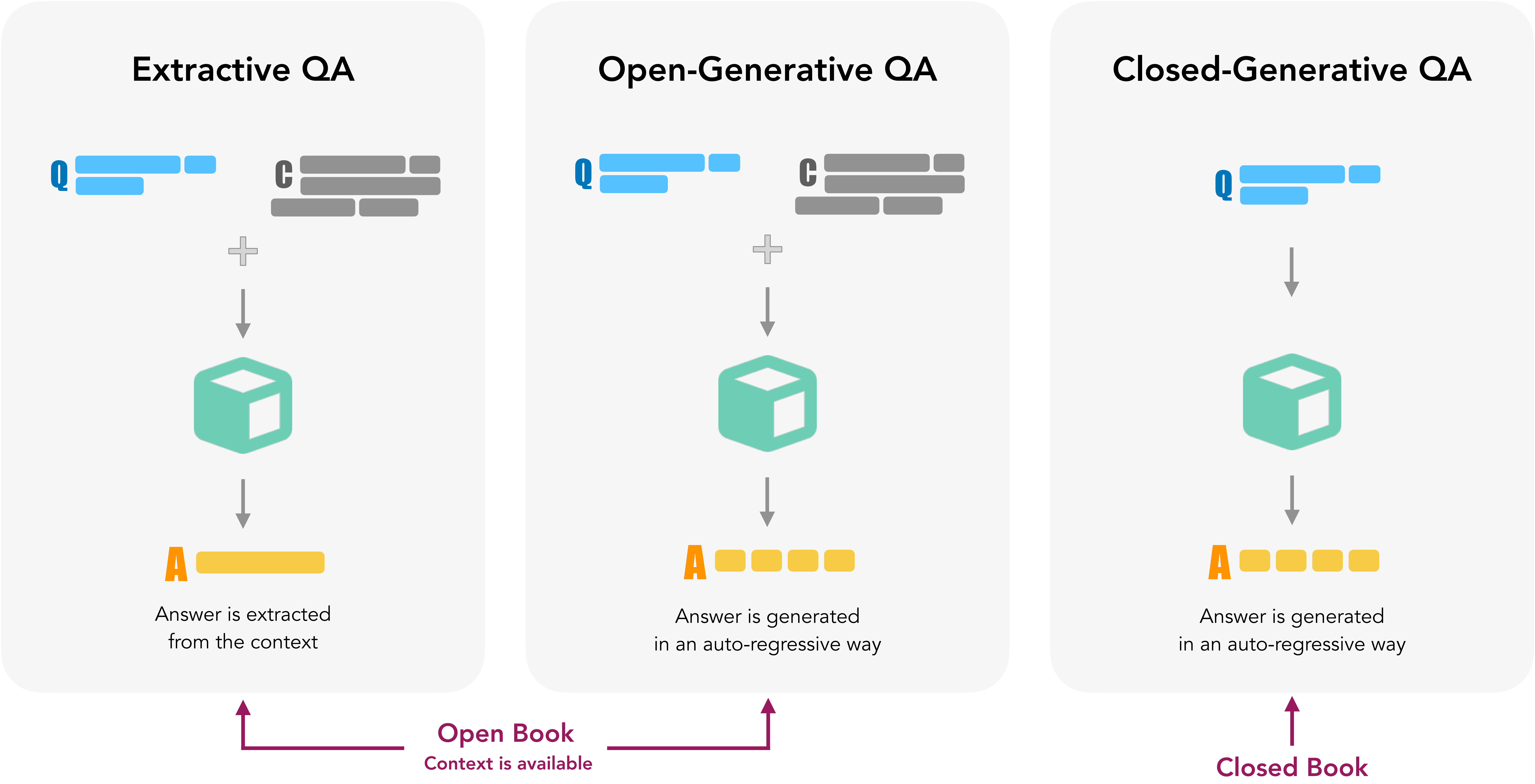

Answer is generated
in an auto-regressive way

## Closed-Generative QA




Answer is generated
in an auto-regressive way

# Types of QA

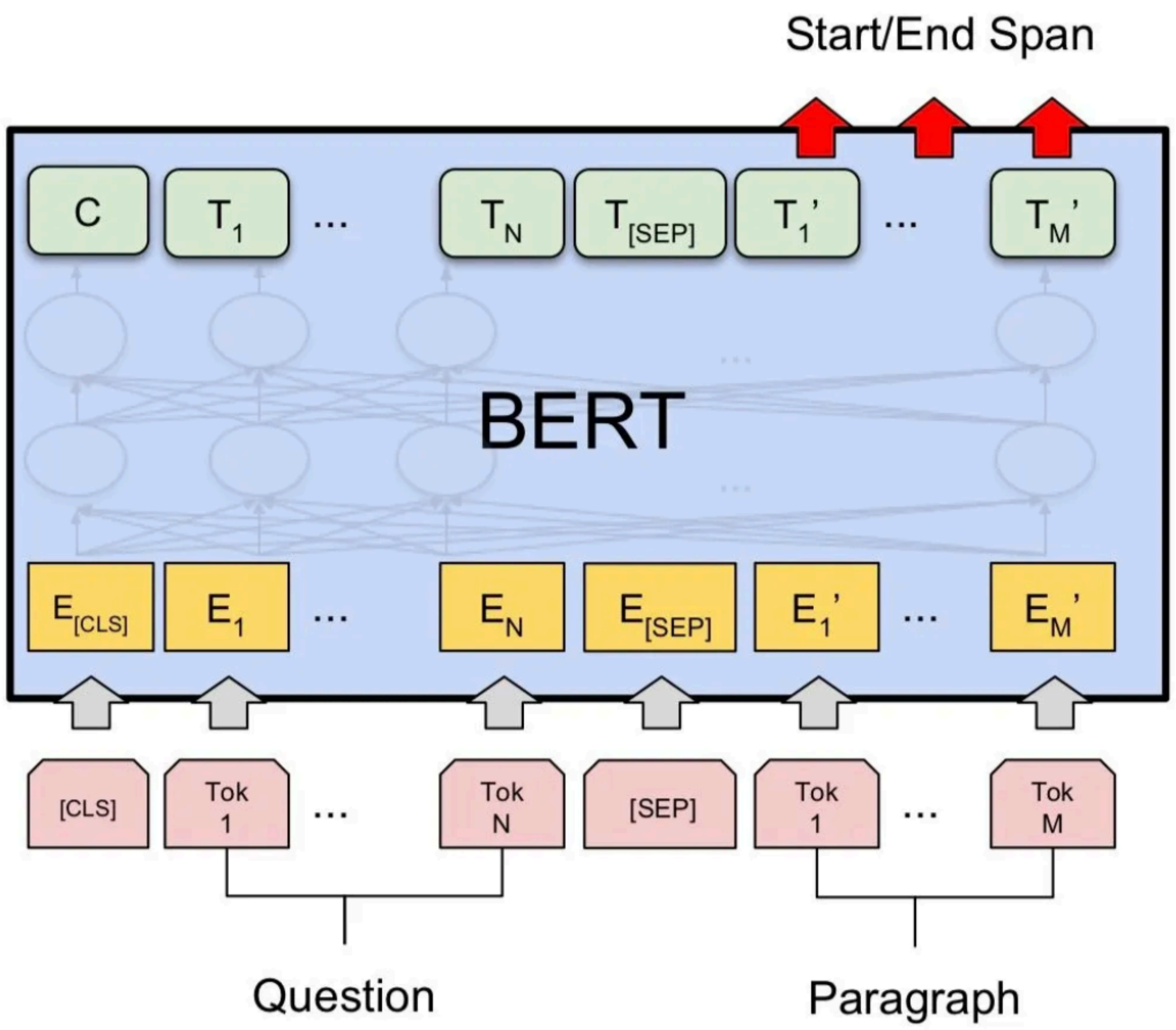


Extractive QA
Q
C
A
Answer is extracted from the context
Open-Generative QA
Q
C
A
Answer is generated in an auto-regressive way
Closed-Generative QA
Q
A
Answer is generated in an auto-regressive way
Open Book
Context is available
Closed Book

# Extractive QA

**Goal:**

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



# Extractive QA

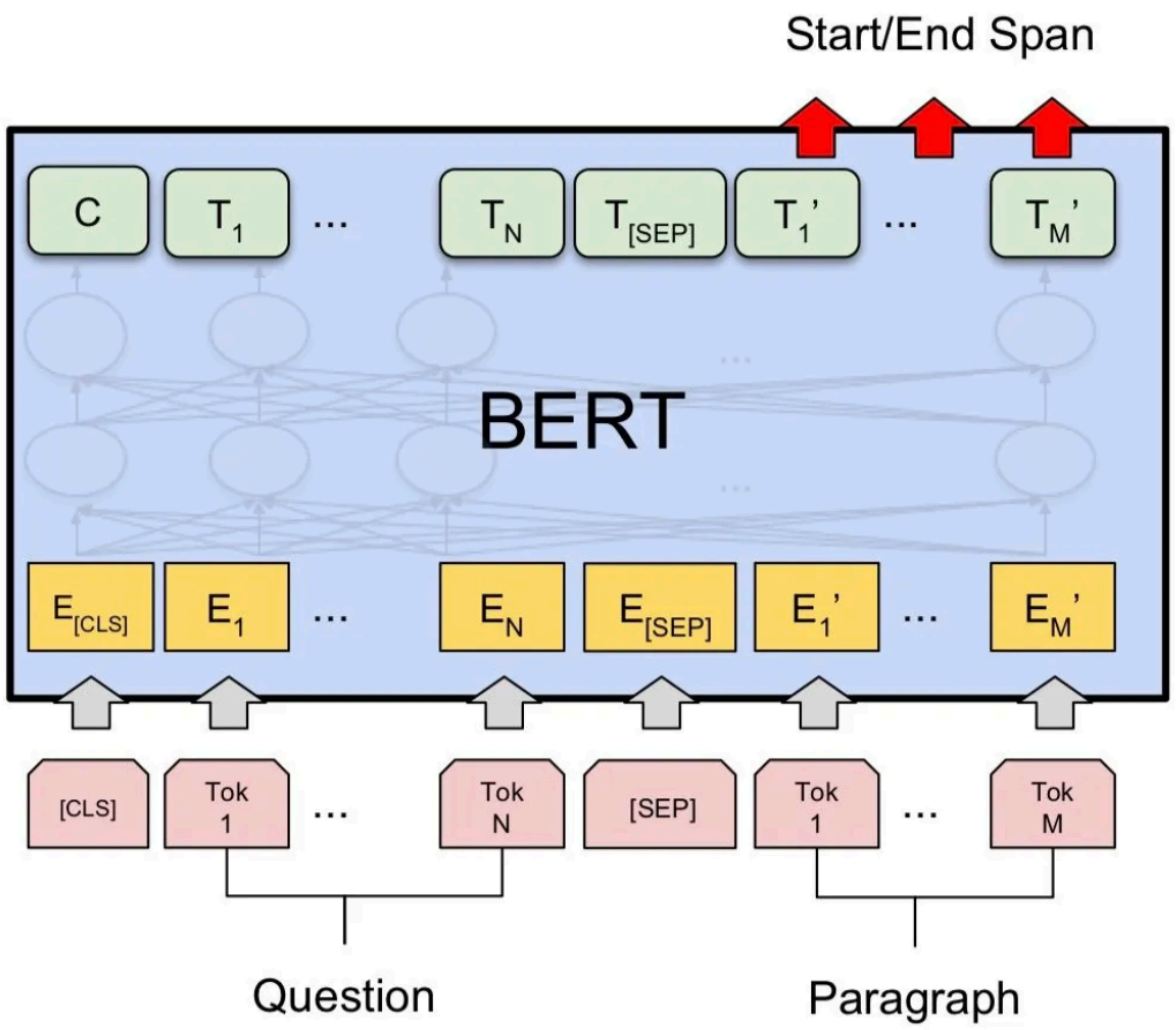
**Goal:**

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

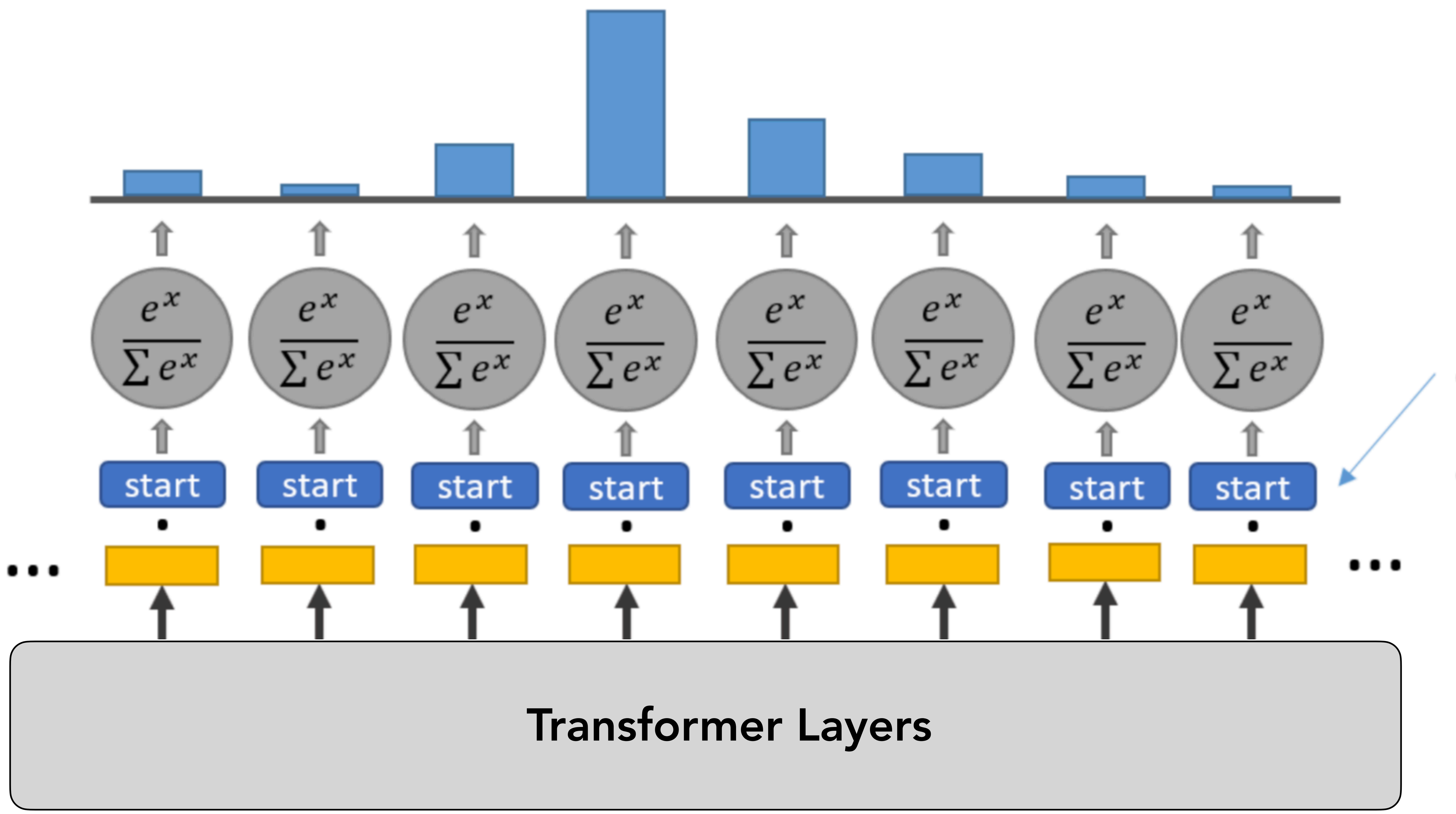
**Input:**

- 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

Rajpurkar et al. (2016)

# 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

| | |
|---|---|
| Train split | 130K |
| Test split | 12K |
| Unanswerable | 50K |

Rajpurkar et al. (2016)

# 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

| | Answer starts |
|---|---|
| 10th and 11th centuries | 94 |
| In the 10th and 11th centuries | 87 |

| | |
|---|---|
| Train split | 130K |
| Test split | 12K |
| Unanswerable | 50K |

Rajpurkar et al. (2016)

# 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

| | Answer starts |
|---|---|
| 10th and 11th centuries | 94 |
| In the 10th and 11th centuries | 87 |

| | |
|---|---|
| Train split | 130K |
| Test split | 12K |
| Unanswerable | 50K |

### 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.

Rajpurkar et al. (2016)

# 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<br>*Stanford University*<br>(Rajpurkar et al. '16) | 82.304 | 91.221 |
| 1<br>Jul 24, 2021 | {ANNA} (single model)<br>*LG AI Research* | **90.622** | **95.719** |
| 2<br>Apr 10, 2020 | LUKE (single model)<br>*Studio Ousia & NAIST & RIKEN AIP*<br>https://arxiv.org/abs/2010.01057 | 90.202 | 95.379 |
| 3<br>May 21, 2019 | XLNet (single model)<br>*Google Brain & CMU* | 89.898 | 95.080 |
| 4<br>Dec 11, 2019 | XLNET-123++ (single model)<br>*MST/EOI*<br>http://tia.today | 89.856 | 94.903 |
| 4<br>Aug 11, 2019 | XLNET-123 (single model)<br>*MST/EOI* | 89.646 | 94.930 |
| 5<br>Jul 21, 2019 | SpanBERT (single model)<br>*FAIR & UW* | 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<br>*Stanford University*<br>(Rajpurkar & Jia et al. '18) | 86.831 | 89.452 |
| 1<br>Jun 04, 2021 | IE-Net (ensemble)<br>*RICOH_SRCB_DML* | **90.939** | **93.214** |
| 2<br>Feb 21, 2021 | FPNet (ensemble)<br>*Ant Service Intelligence Team* | 90.871 | 93.183 |
| 3<br>May 16, 2021 | IE-NetV2 (ensemble)<br>*RICOH_SRCB_DML* | 90.860 | 93.100 |
| 4<br>Apr 06, 2020 | SA-Net on Albert (ensemble)<br>*QIANXIN* | 90.724 | 93.011 |
| 5<br>May 05, 2020 | SA-Net-V2 (ensemble)<br>*QIANXIN* | 90.679 | 92.948 |
| 5<br>Apr 05, 2020 | Retro-Reader (ensemble)<br>*Shanghai Jiao Tong University*<br>http://arxiv.org/abs/2001.09694 | 90.578 | 92.978 |
| 5<br>Feb 05, 2021 | FPNet (ensemble)<br>*YuYang* | 90.600 | 92.899 |

https://rajpurkar.github.io/SQuAD-explorer/

# 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**

*Jia and Liang, 2017*

# 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**

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

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

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

**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

## Cons of Extractive:

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

## Pros of Generative:

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

# 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.)

## Pros of Generative:

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

## Cons of Generative:

- Hallucinations / lexical repetitions
- Grammar mistakes

# 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.)

## Pros of Generative:

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

## 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)***

"Who is the president of France?"

| Golden answer | Predicted answers | EM | F1 |
|---|---|---|---|
| Emmanuel Macron | Emmanuel Macron | ✅ | ✅ |
| | Emmanuel Jean-Michel Frédéric Macron | ❌ | ✅ |

## 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.

# 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 | ❌ | ✅ |

## 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.

**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.

Kwiatkowski et al. (2019)

# 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.)

Kwiatkowski et al. (2019)

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

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

**<u>Question:</u> 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.