

Incorporating Extra-linguistic Contexts and Entity Knowledge into NLP

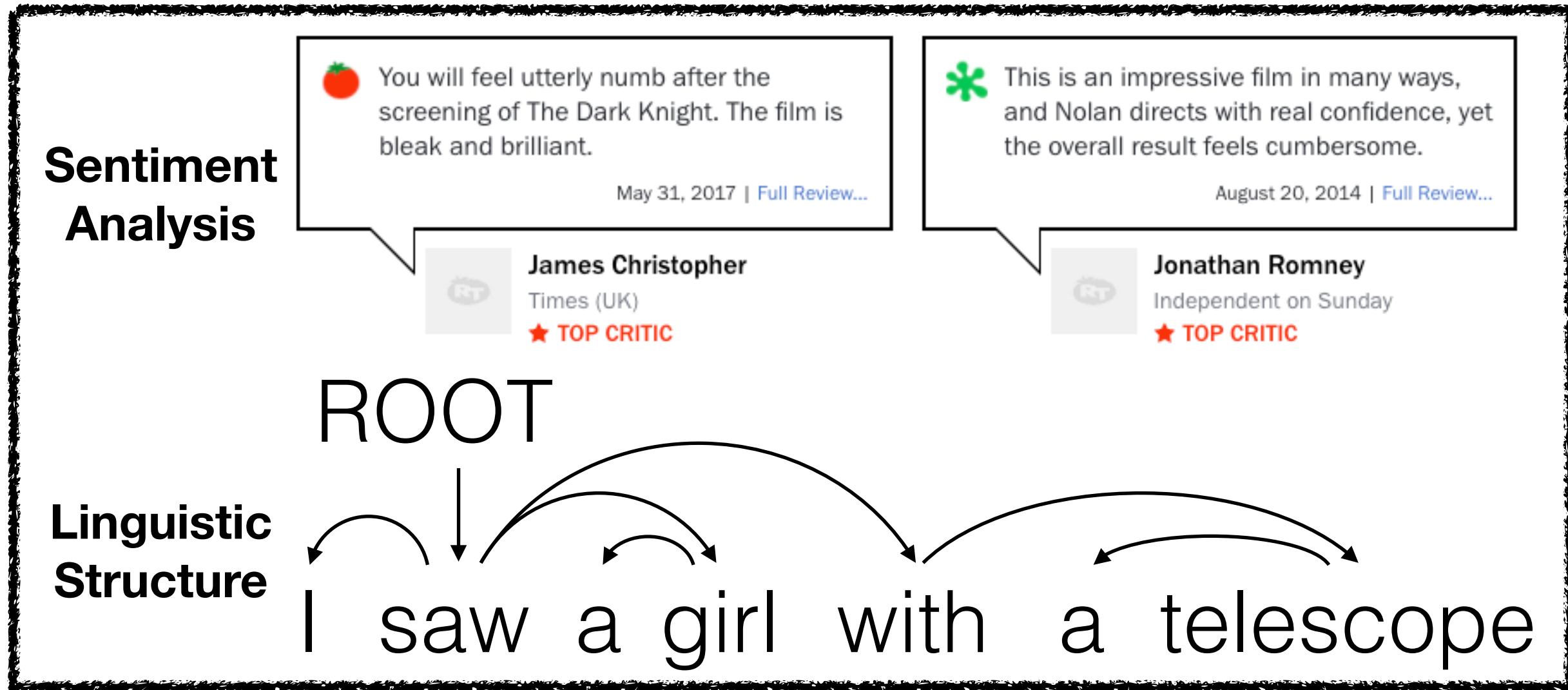
Eunsol Choi



October 7th, 2021

AKBC Workshop, Unstructured and Structured KB

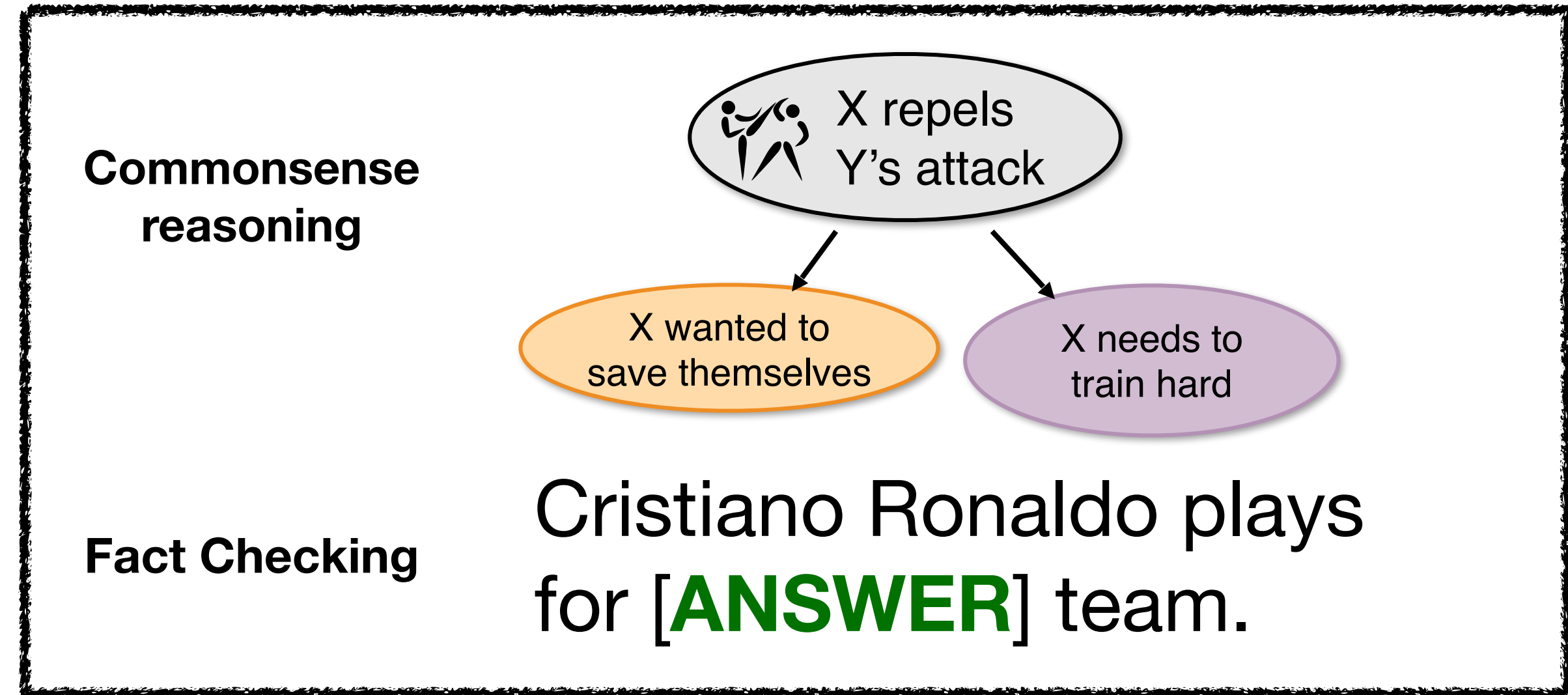
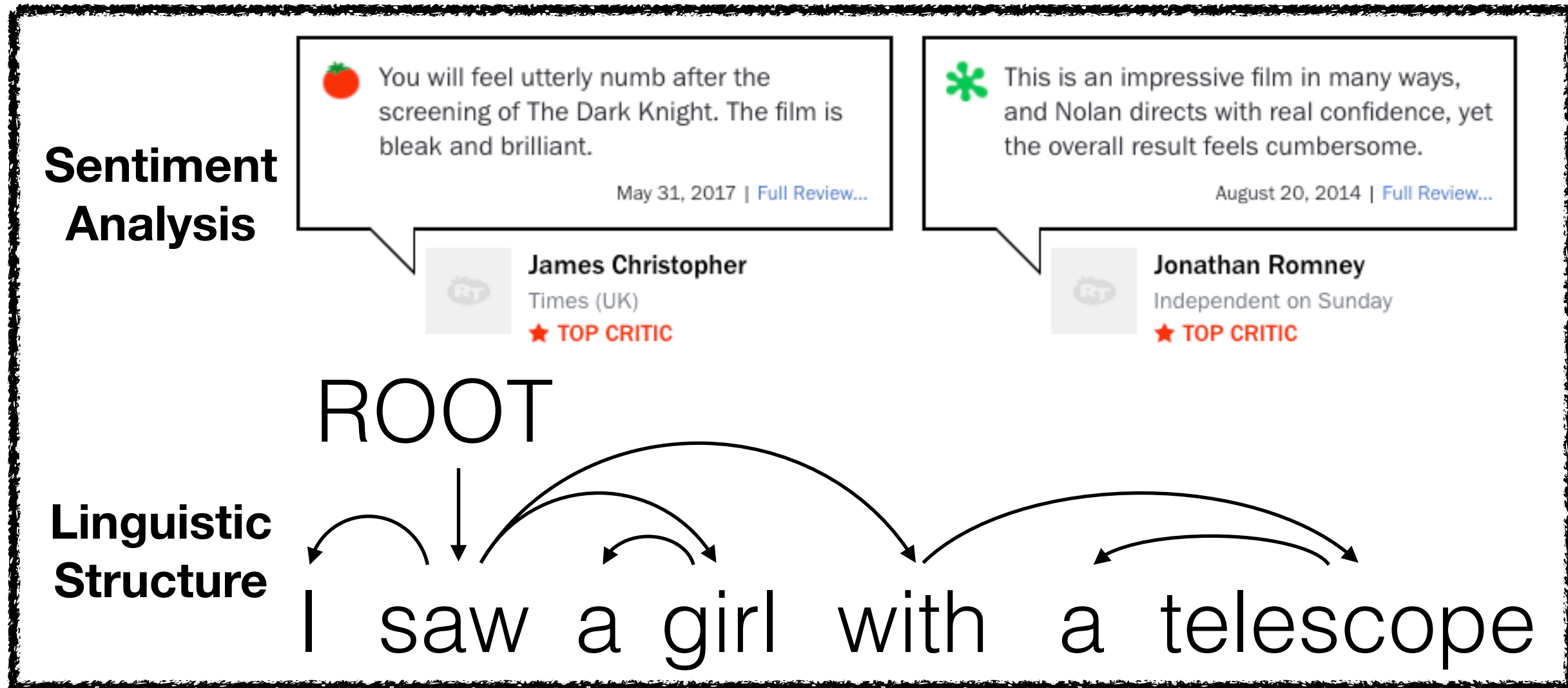
How NLP benchmarks are changing



Past

- Tasks focus on lexical, linguistic knowledge
- All the necessary context is provided

How NLP benchmarks are changing



Past

- Tasks focus on lexical, linguistic knowledge
- All the necessary context is provided

Today

- Further requires factual knowledge and reasoning based on common sense

How NLP benchmarks are changing: QA

Question : What shift happened in animal regulation in 1963 in U.S?

Document Context: ...Whereas the Lacey Act dealt with game animal management and market commerce species, a major shift in focus occurred by 1963 to habitat preservation instead of take regulations. A provision was added by Congress in the Land and Water Conservation Fund Act of...

Answer is span in the provided document

Question : Where was the last Winter Olympic Games held?



WIKIPEDIA
The Free Encyclopedia



Answer: Sochi

Past

- All necessary context is provided in the evidence document

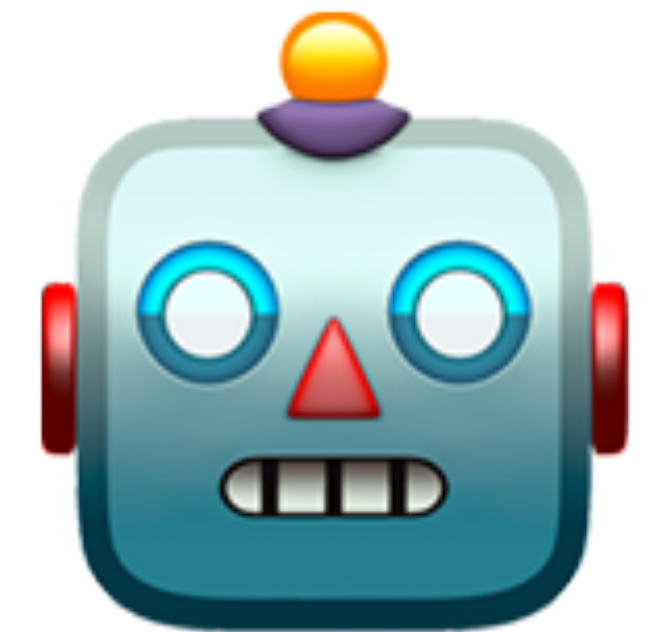
Today

- The model is required to find answer in an unconstrained setting

Open Retrieval QA: missing contexts



Cristiano Ronaldo plays
for [ANSWER] team.

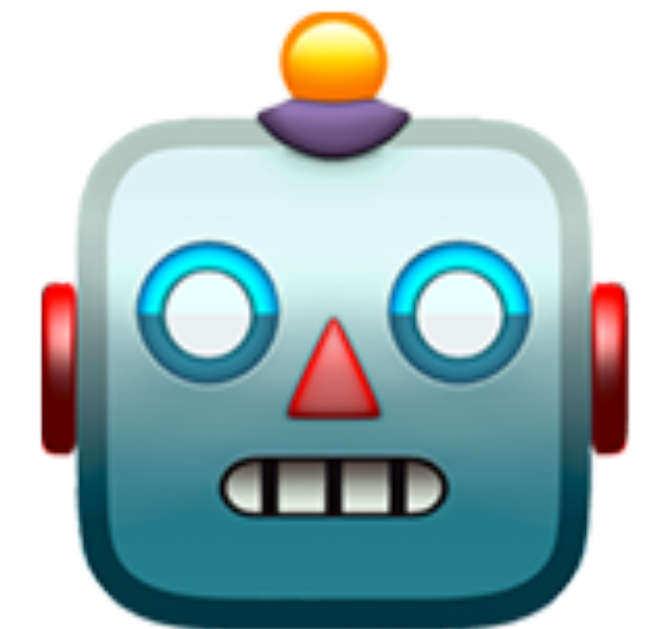


Open Retrieval QA: missing contexts



Cristiano Ronaldo plays for [ANSWER] team.

Cristiano Ronaldo		
Youth career		
1992–1995	Andorinha	
1995–1997	Nacional	
1997–2002	Sporting CP	
Senior career*		
Years	Team	Apps (Gls)
2002–2003	Sporting CP B	2 (0)
2002–2003	Sporting CP	25 (3)
2003–2009	Manchester United	196 (84)
2009–2018	Real Madrid	292 (311)
2018–2021	Juventus	98 (81)
2021–	Manchester United	4 (3)



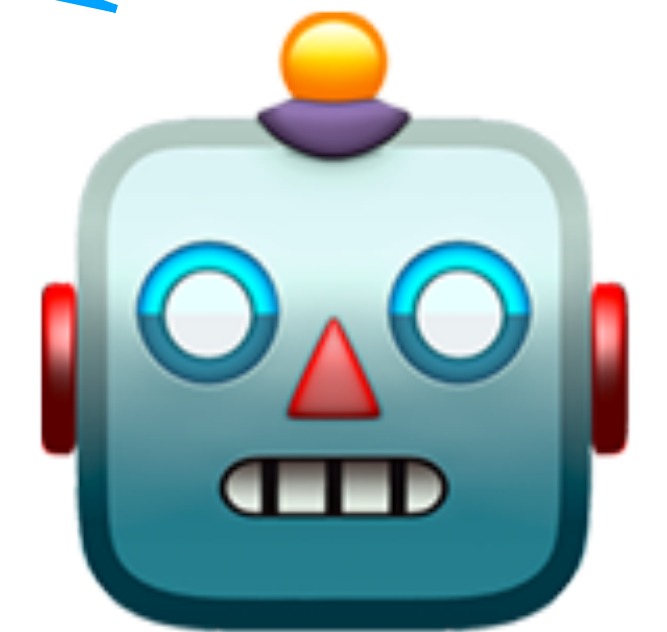
Open Retrieval QA: missing contexts



Cristiano Ronaldo plays for [ANSWER] team.

When are you asking this question?

 temporal context



Cristiano Ronaldo

Youth career

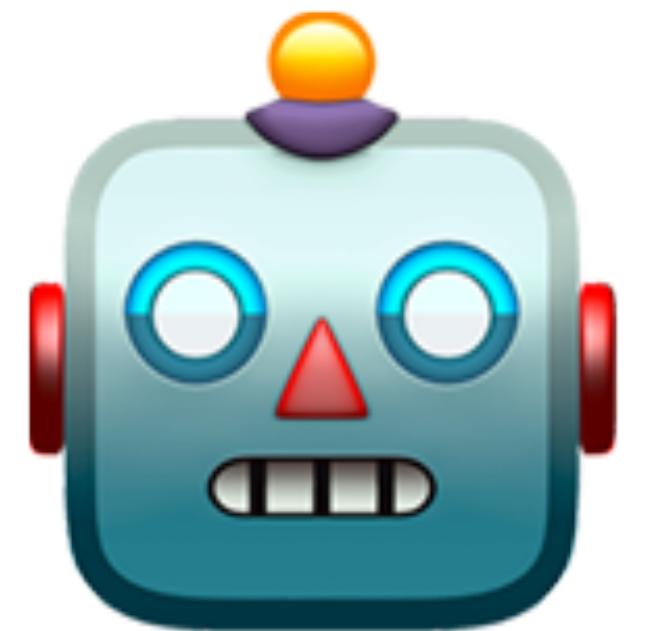
1992–1995	Andorinha
1995–1997	Nacional
1997–2002	Sporting CP

Senior career*

Years	Team	Apps	(Gls)
2002–2003	Sporting CP B	2	(0)
2002–2003	Sporting CP	25	(3)
2003–2009	Manchester United	196	(84)
2009–2018	Real Madrid	292	(311)
2018–2021	Juventus	98	(81)
2021–	Manchester United	4	(3)

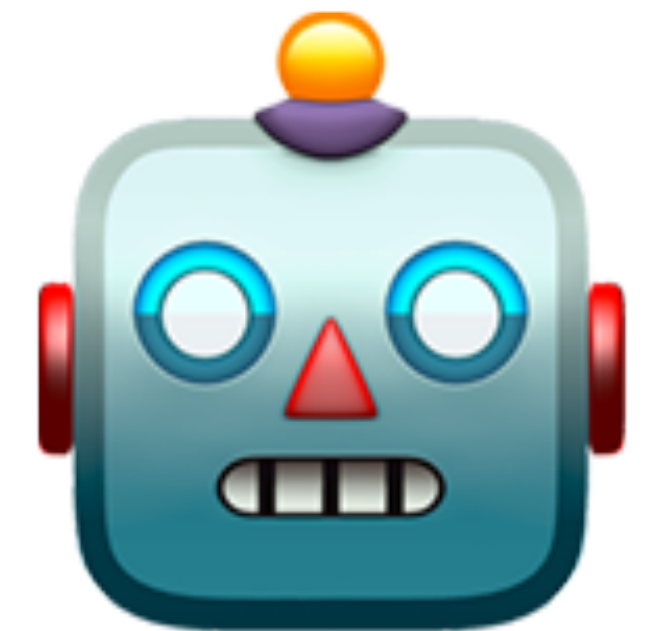
Open Retrieval QA: missing contexts

How many gold medals did we win in the Winter Olympics 2018?



Open Retrieval QA: missing contexts

How many gold medals did we win in the Winter Olympics 2018?



2018 Winter Olympics Medal Table

Rank ↕	NOC ▲	Gold ↕
? 23	 Australia (AUS)	0
? 10	 Austria (AUT)	5
? 15	 Belarus (BLR)	2
? 25	 Belgium (BEL)	0

Open Retrieval QA: missing contexts



How many gold medals did we win in the Winter Olympics 2018?

Where are you asking this question?

 geographical context

2018 Winter Olympics Medal Table



Rank ↕	NOC	Gold ↕
? 23	 Australia (AUS)	0
? 10	 Austria (AUT)	5
? 15	 Belarus (BLR)	2
? 25	 Belgium (BEL)	0

Incorporating extra-linguistic contexts into QA



This Talk

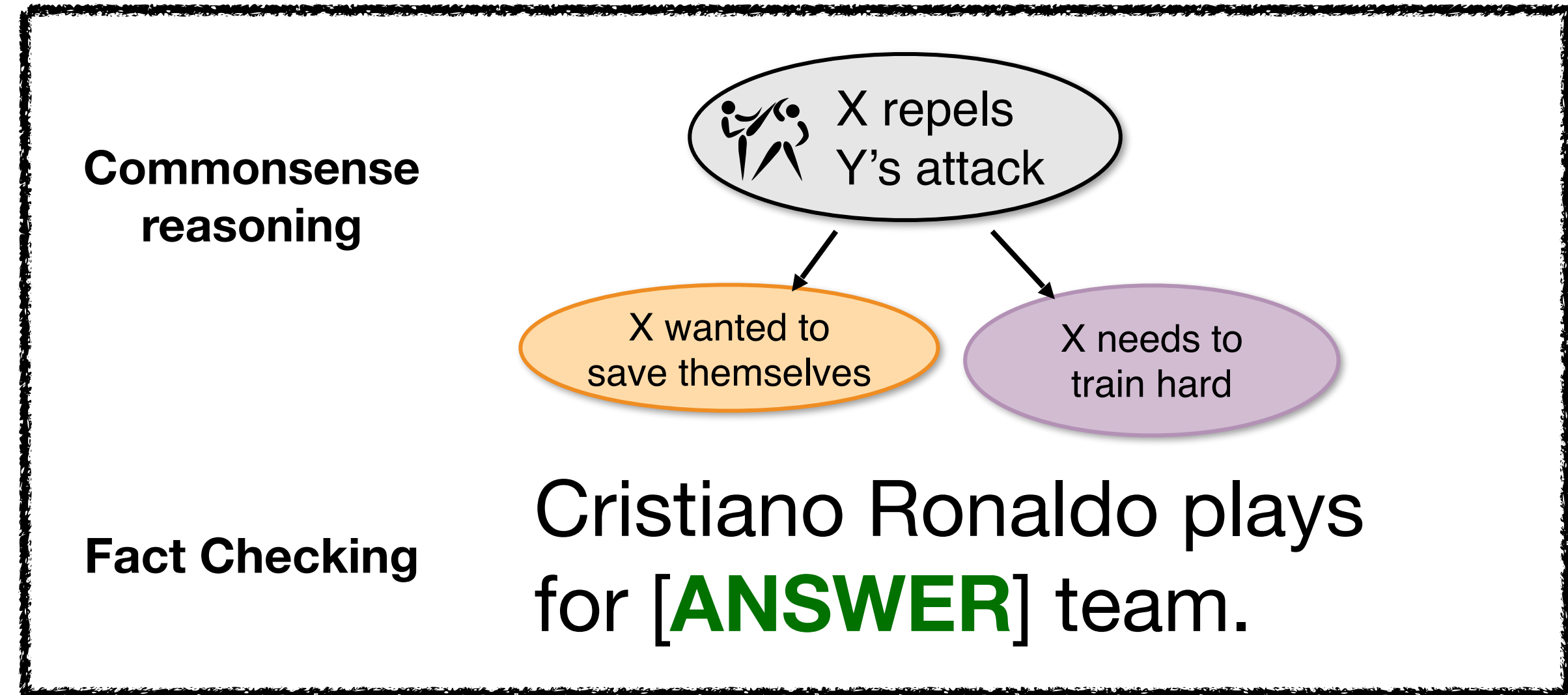
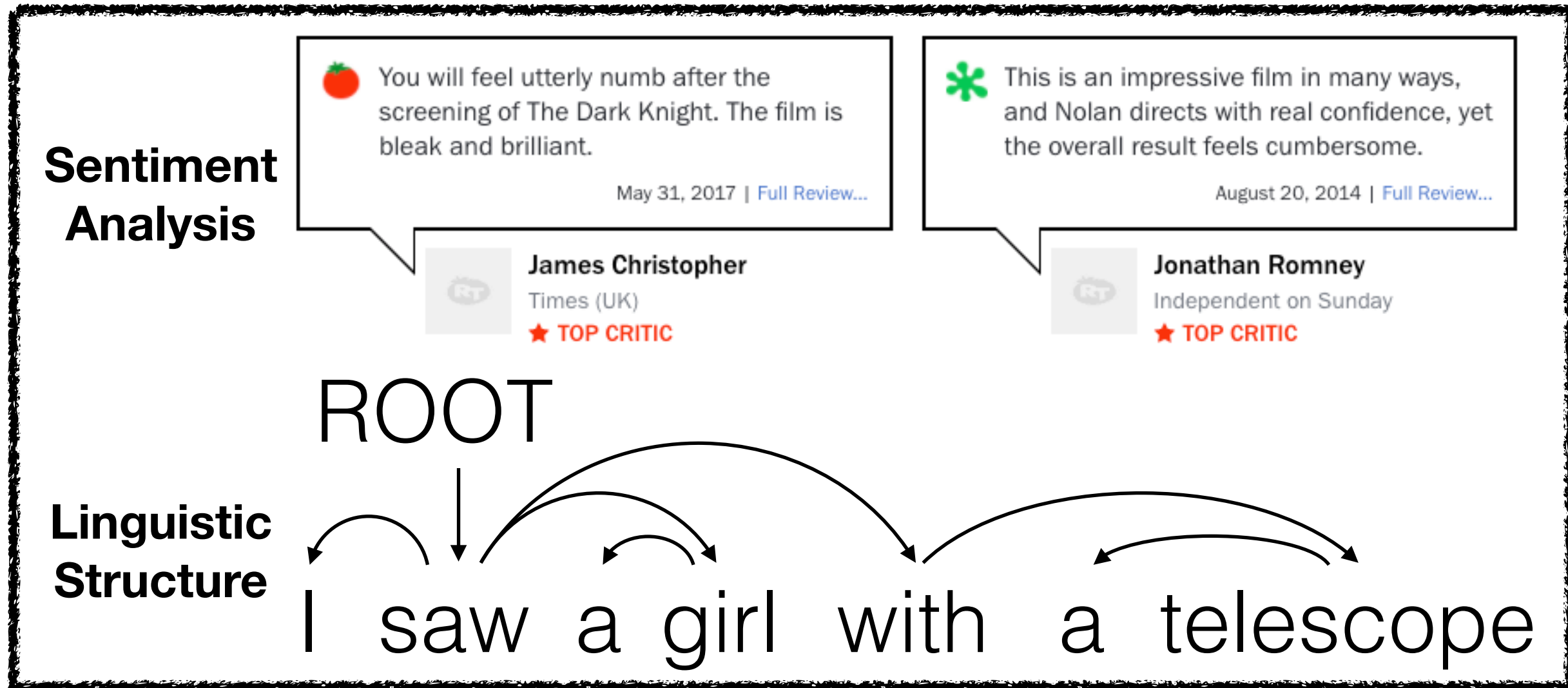
- Incorporating two extra-linguistic contexts (🕒 temporal and 🌐 geographical) into open retrieval question answering

SituatedQA: Incorporating Extra-Linguistic Contexts into QA [EMNLP 2021]

Michael J.Q. Zhang and Eunsol Choi



How NLP benchmarks are changing



Past

- Tasks focus on lexical, linguistic knowledge
- All the necessary context is provided

Today

- Further requires factual knowledge and reasoning based on common sense

Two types of knowledge rich tasks

Fact Verification [Vlachos and Riedel, 2014, Thorne et al., 2018]



WIKIPEDIA

- Simple facts on real world entities

Claim: There exists a producer and an actor called Simon Pegg

Supported

Claim: Mary McGee was the first woman to compete in the Baja 1000, between 1971 and 1979.

Refuted

Two types of knowledge rich tasks

Fact Verification [Vlachos and Riedel, 2014, Thorne et al., 2018]

- Simple facts on real world entities



Claim: There exists a producer and an actor called Simon Pegg

Supported

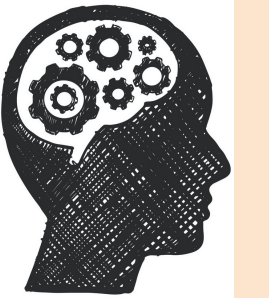
Claim: Mary McGee was the first woman to compete in the Baja 1000, between 1971 and 1979.

Refuted

Commonsense Reasoning

[Levesque et al., 2011, Talmor et al., 2019, 2021]

- Reasoning on fictional / general entities



*Where on a **river** can you hold a cup upright to catch water on a sunny day?*

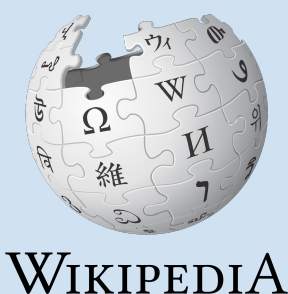
✓ **waterfall**, ✗ **bridge**, ✗ **valley**, ✗ **pebble**, ✗ **mountain**

*Where can I stand on a **river** to see water falling without getting wet?*

✗ **waterfall**, ✓ **bridge**, ✗ **valley**, ✗ **stream**, ✗ **bottom**

Our work: reasoning based on facts about entities

Fact Verification [Vlachos and Riedel, 2014, Thorne et al., 2018]



- Simple facts on real world entities

Claim: There exists a producer and an actor called Simon Pegg

Supported

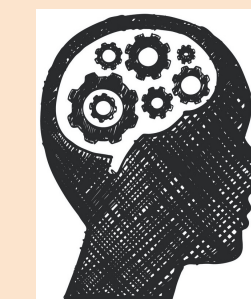
Claim: Mary McGee was the first woman to compete in the Baja 1000, between 1971 and 1979.

Refuted

Commonsense Reasoning

[Levesque et al., 2011, Talmor et al., 2019, 2021]

- Reasoning on fictional / general entities



*Where on a **river** can you hold a cup upright to catch water on a sunny day?*

✓ **waterfall**, ✗ **bridge**, ✗ **valley**, ✗ **pebble**, ✗ **mountain**

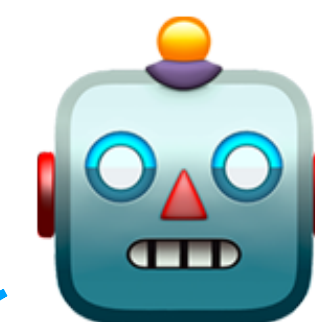
*Where can I stand on a **river** to see water falling without getting wet?*

✗ **waterfall**, ✓ **bridge**, ✗ **valley**, ✗ **stream**, ✗ **bottom**

Claim: **Harry Potter** can teach classes on how to fly on a broomstick.

True?

False?



Our work: reasoning based on facts about entities

Fact Verification [Vlachos and Riedel, 2014, Thorne et al., 2018]



- Simple facts on real world entities

Claim: There exists a producer and an actor called Simon Pegg

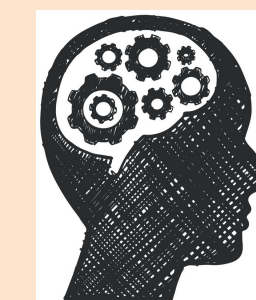
Supported

Claim: Mary McGee was the first woman to compete in the Baja 1000, between 1971 and 1979.

Refuted

Commonsense Reasoning

[Levesque et al., 2011, Talmor et al., 2019, 2021]



- Reasoning on fictional / general entities

*Where on a **river** can you hold a cup upright to catch water on a sunny day?*

✓ **waterfall**, ✗ **bridge**, ✗ **valley**, ✗ **pebble**, ✗ **mountain**

*Where can I stand on a **river** to see water falling without getting wet?*

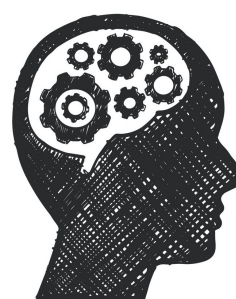
✗ **waterfall**, ✓ **bridge**, ✗ **valley**, ✗ **stream**, ✗ **bottom**

Claim: **Harry Potter** can teach classes on how to fly on a broomstick.



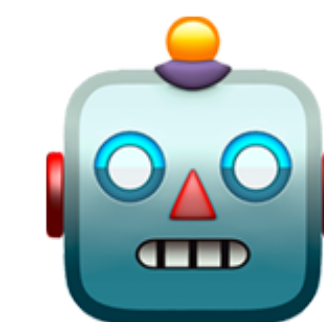
Harry Potter is a wizard ...
He plays Quidditch while riding on a broomstick.

+



Someone who's good at something can teach it.

True

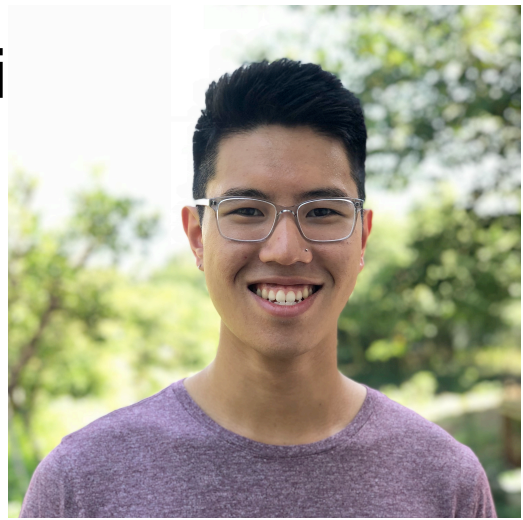


This Talk

- Incorporating two extra-linguistic contexts (🕒 temporal and 🌍 geographical) into open retrieval question answering
- Presenting a benchmark which evaluates model's reasoning ability anchored at entity information

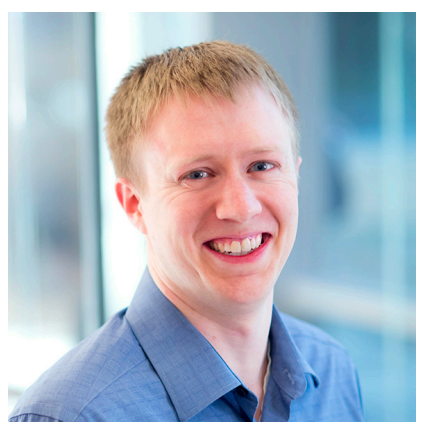
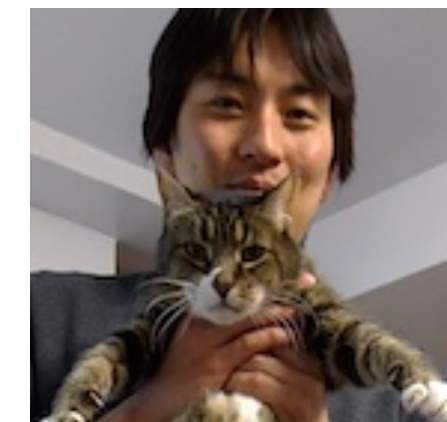
SituatedQA: Incorporating Extra-Linguistic Contexts into QA [EMNLP 2021]

Michael J.Q. Zhang and Eunsol Choi



CREAK: A Dataset for Commonsense Reasoning over Entity Knowledge [In submission]

Yasumasa Onoe, Michael J.Q. Zhang, Eunsol Choi, Greg Durrett



How commonly questions depend on contexts?



Cristiano Ronaldo plays for [ANSWER] team.

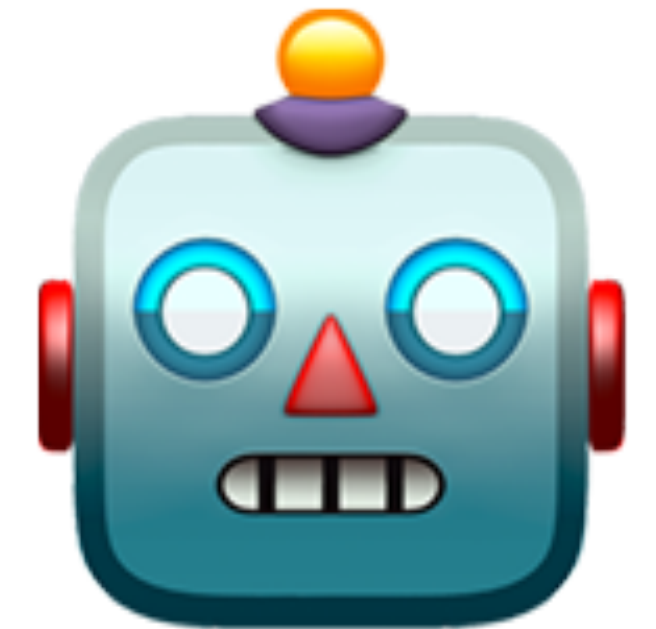
How many gold medals did we win in the Winter Olympics 2018?

When are you asking this question?

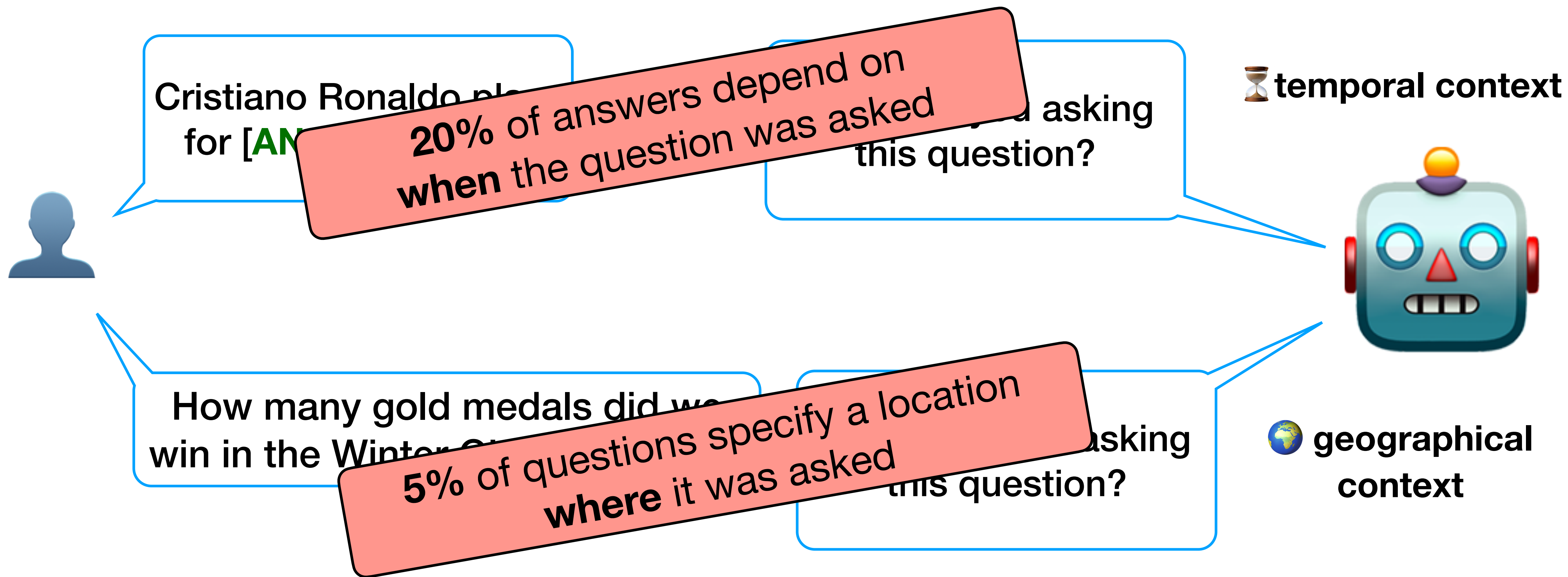
Where are you asking this question?

 temporal context

 geographical context



How commonly questions depend on contexts?



SituatedQA: Our Goals

- 1. Create QA systems that can produce the appropriate answer for a given temporal or geographical context**
- 2. Analyze existing benchmarks and models how they handle such extra linguistic contexts**

The SituatedQA Task: (question, context) → answer



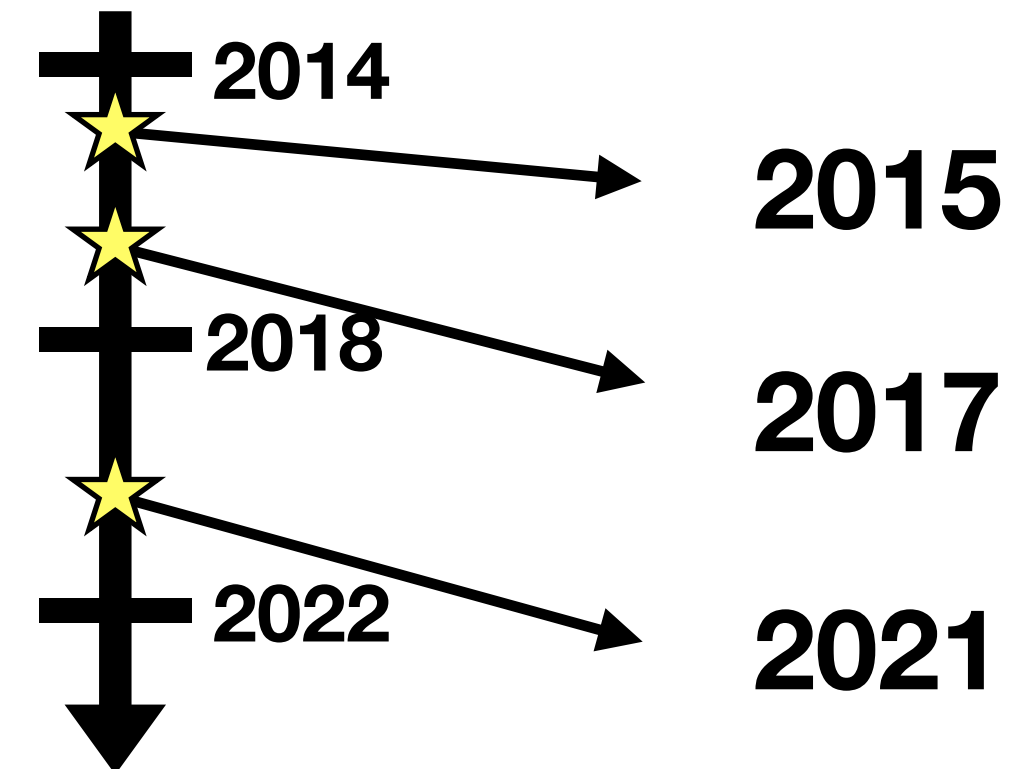
Question

Where was the last Winter Olympic Games held?



,

Temporal Context



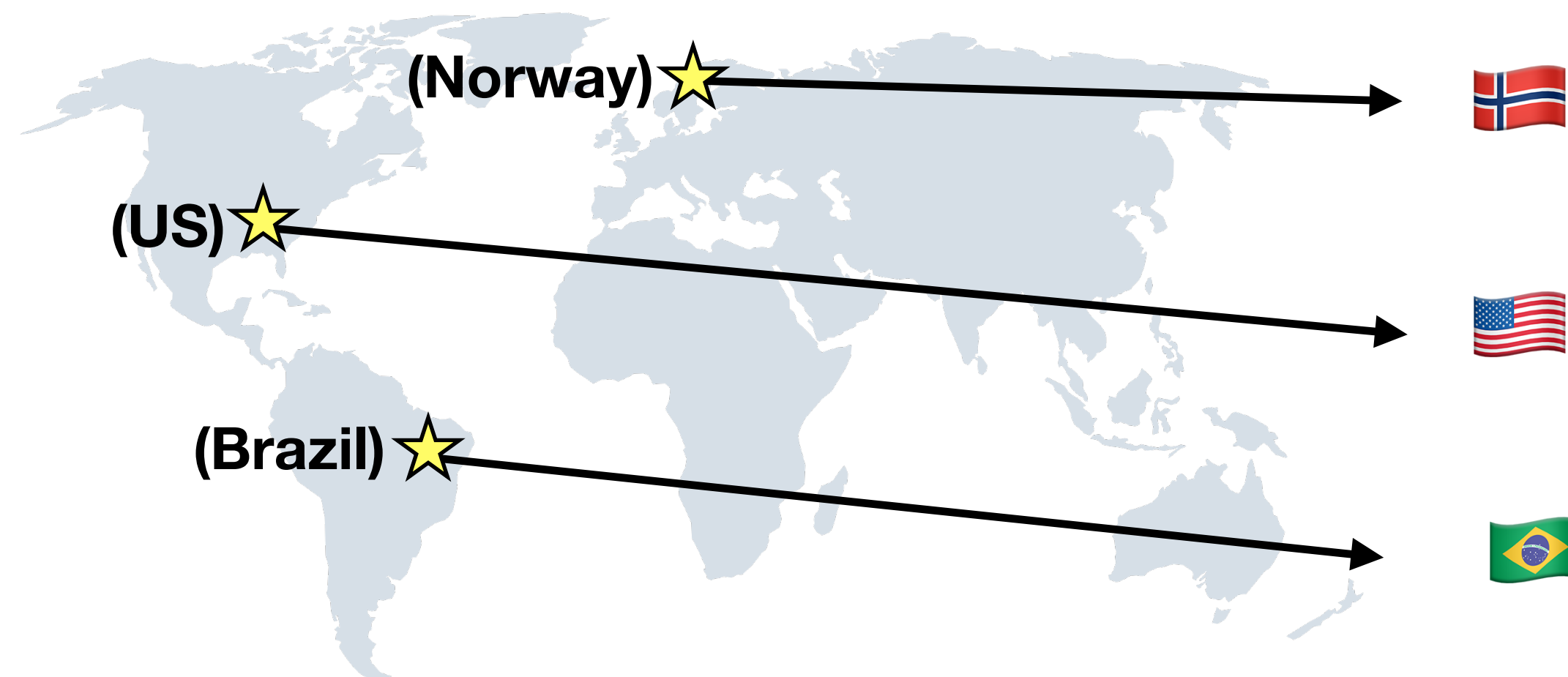
Question

How many gold medals did we win in the Winter Olympics 2018?



,

Geographical Context



The SituatedQA Task: (question, context) → answer



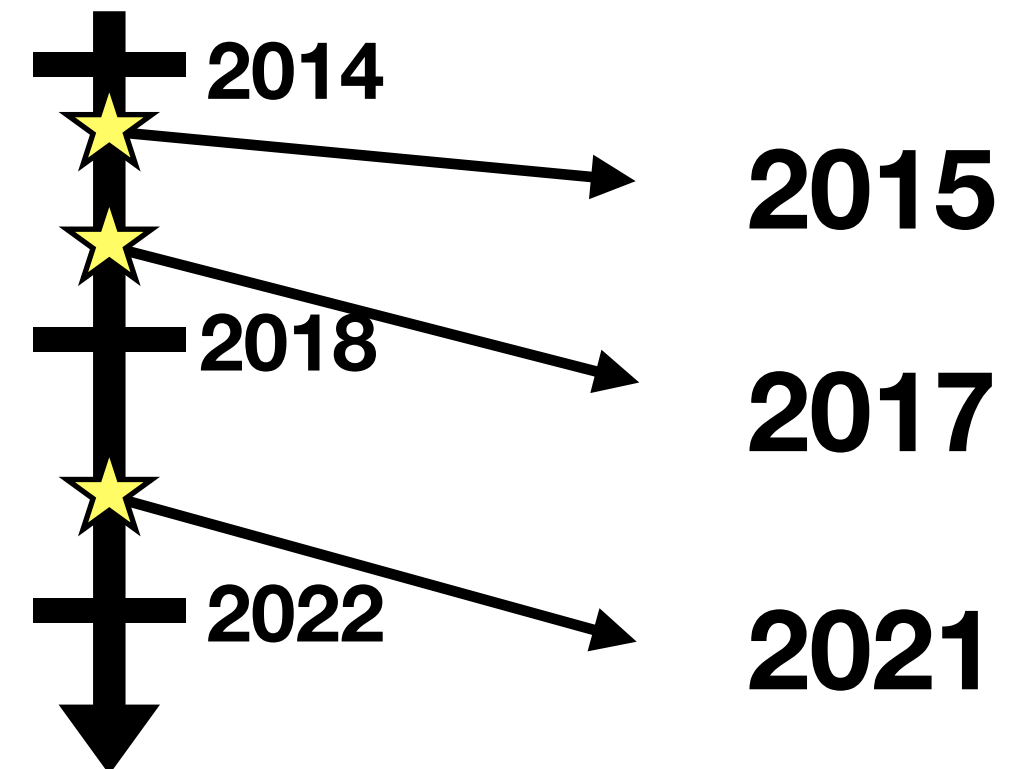
Question

Where was the last Winter Olympic Games held?



,

Temporal Context



Answer

Sochi

Sochi

Pyeongchang



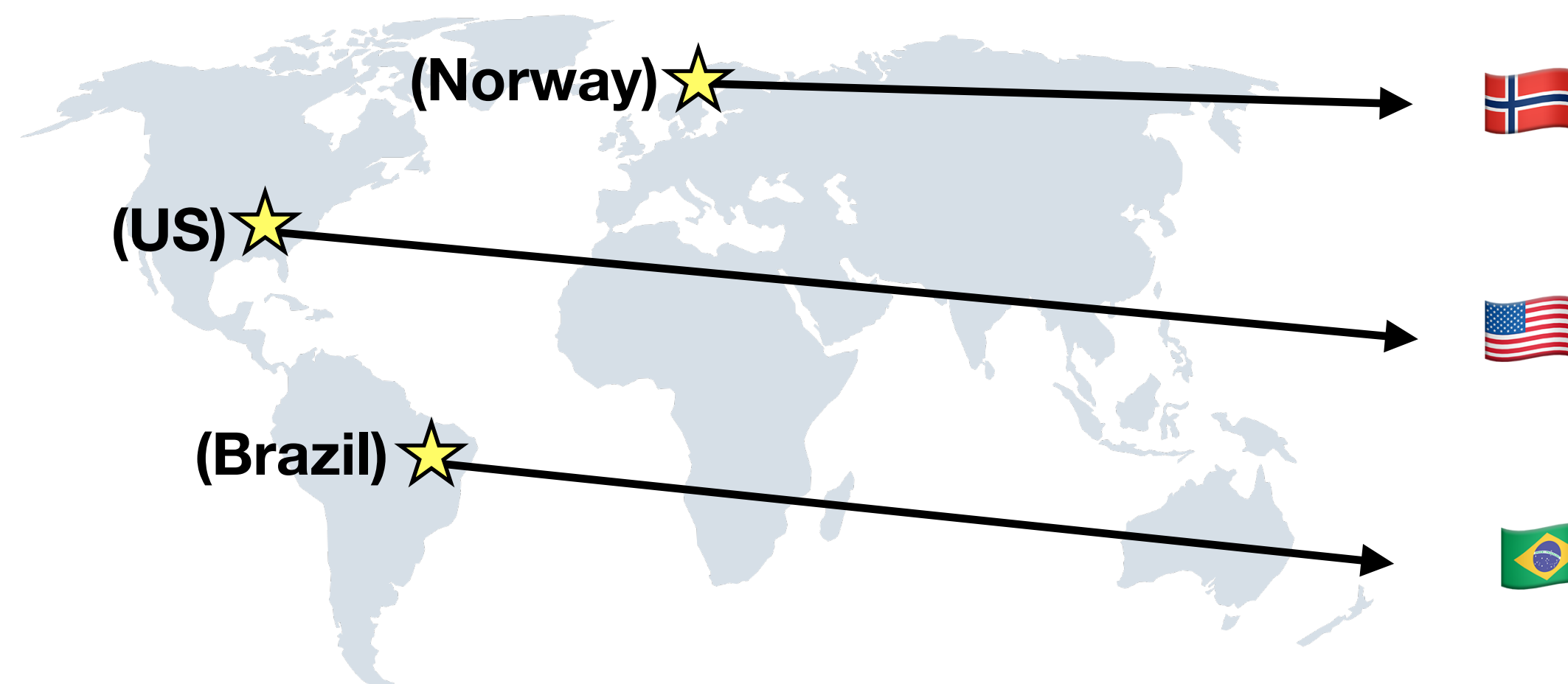
Question

How many gold medals did we win in the Winter Olympics 2018?



,

Geographical Context



Answer

14

9

0

Dataset: Sourcing Questions for SituatedQA

Temporal Dependence in QA

Temporally dependent questions are *abundant* in existing datasets

- Annotators provide the correct answer at the time of annotation
- We find that **10-30%** of examples in TyDi-QA (English Only), WebQuestions, MS MARCO, NQ-Open are **temporally-dependent**

Dataset: Sourcing Questions for SituatedQA

Temporal Dependence in QA

Temporally dependent questions are *abundant* in existing datasets

- Annotators provide the correct answer at the time of annotation
- We find that **10-30%** of examples in TyDi-QA (English Only), WebQuestions, MS MARCO, NQ-Open are **temporally-dependent**

Geographically Dependence in QA

Geographically dependent questions are *absent* in existing datasets

- There is no natural “correct” geographical context for annotators

Dataset: Sourcing Questions for SituatedQA

Temporal Dependence in QA

Temporally dependent questions are *abundant* in existing datasets

- Annotators provide the correct answer at the time of annotation
- We find that **10-30%** of examples in TyDi-QA (English Only), WebQuestions, MS MARCO, NQ-Open are **temporally-dependent**

Geographically Dependence in QA

Geographically dependent questions are *absent* in existing datasets

- There is no natural “correct” geographical context for annotators
- **To create such questions**, we edit existing questions that specify a location

Who is the president of the US?



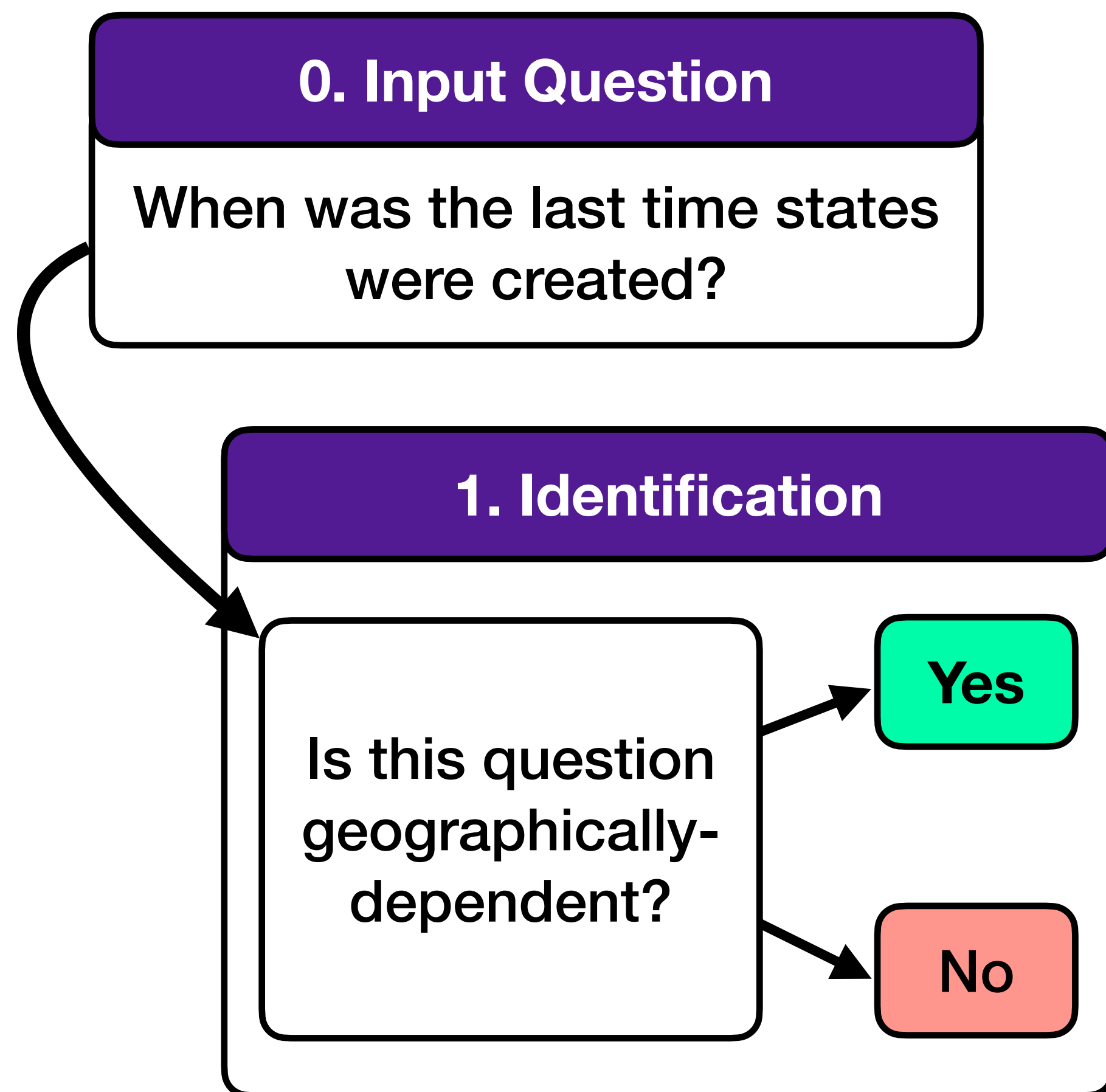
Who is the president?

Dataset: Geographical SituatedQA

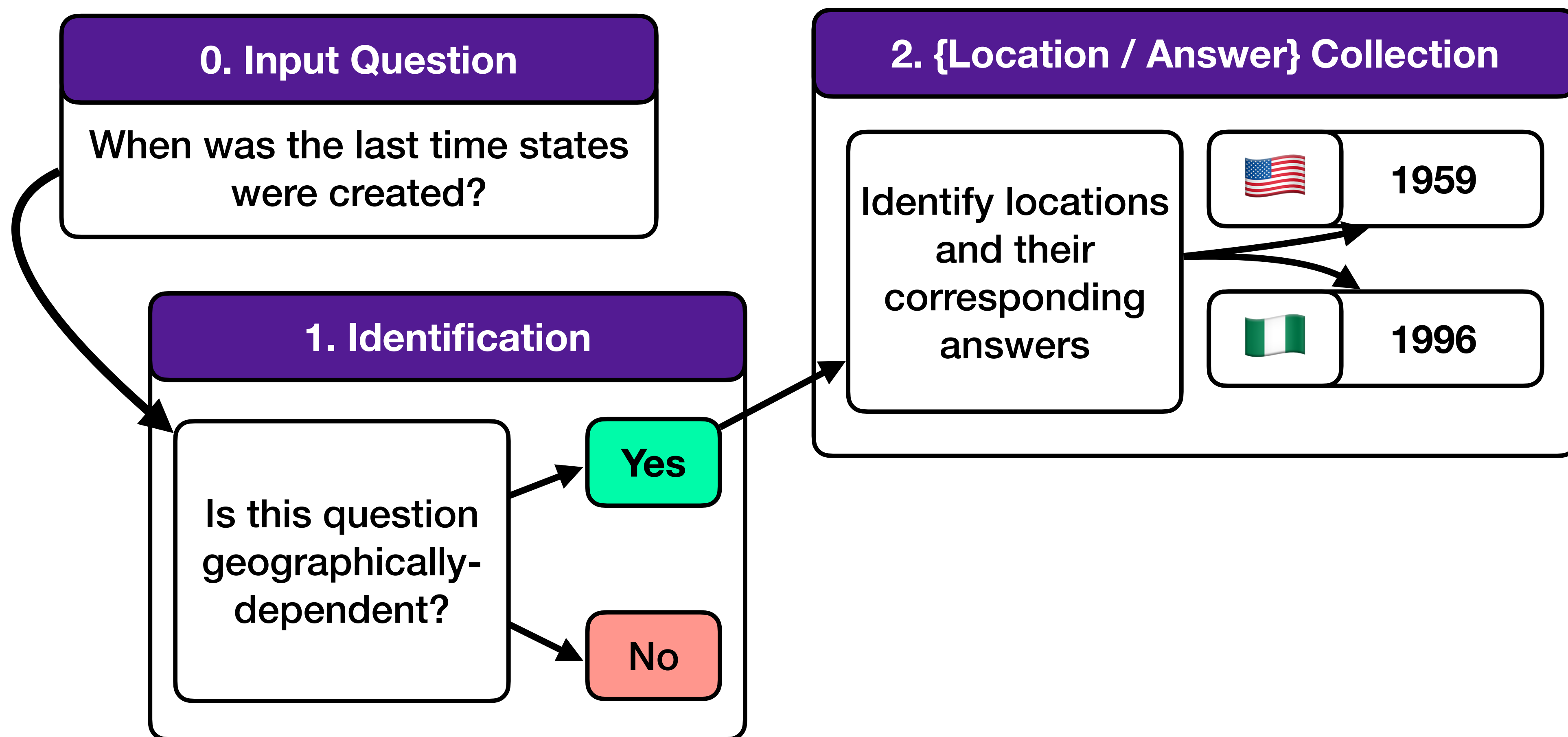
0. Input Question

When was the last time states were created?

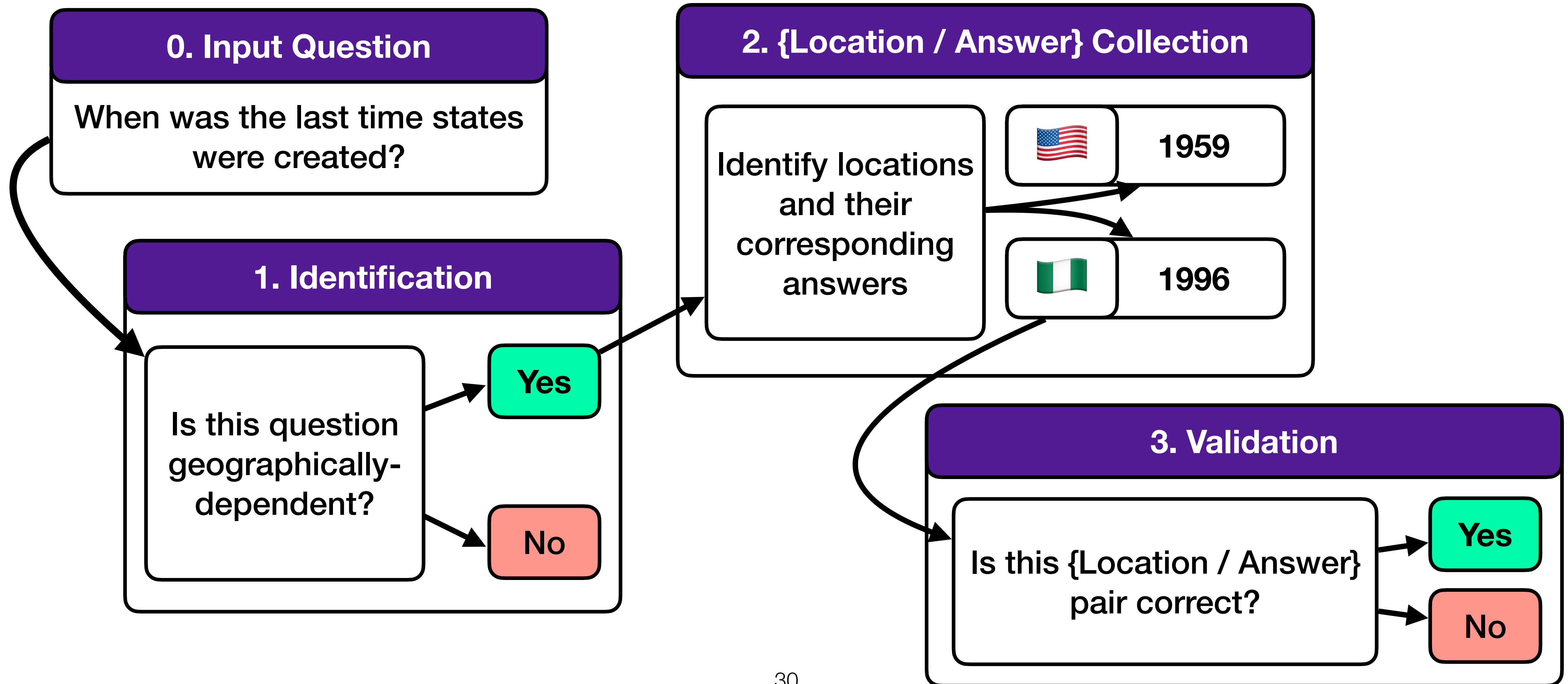
Dataset: Geographical SituatedQA



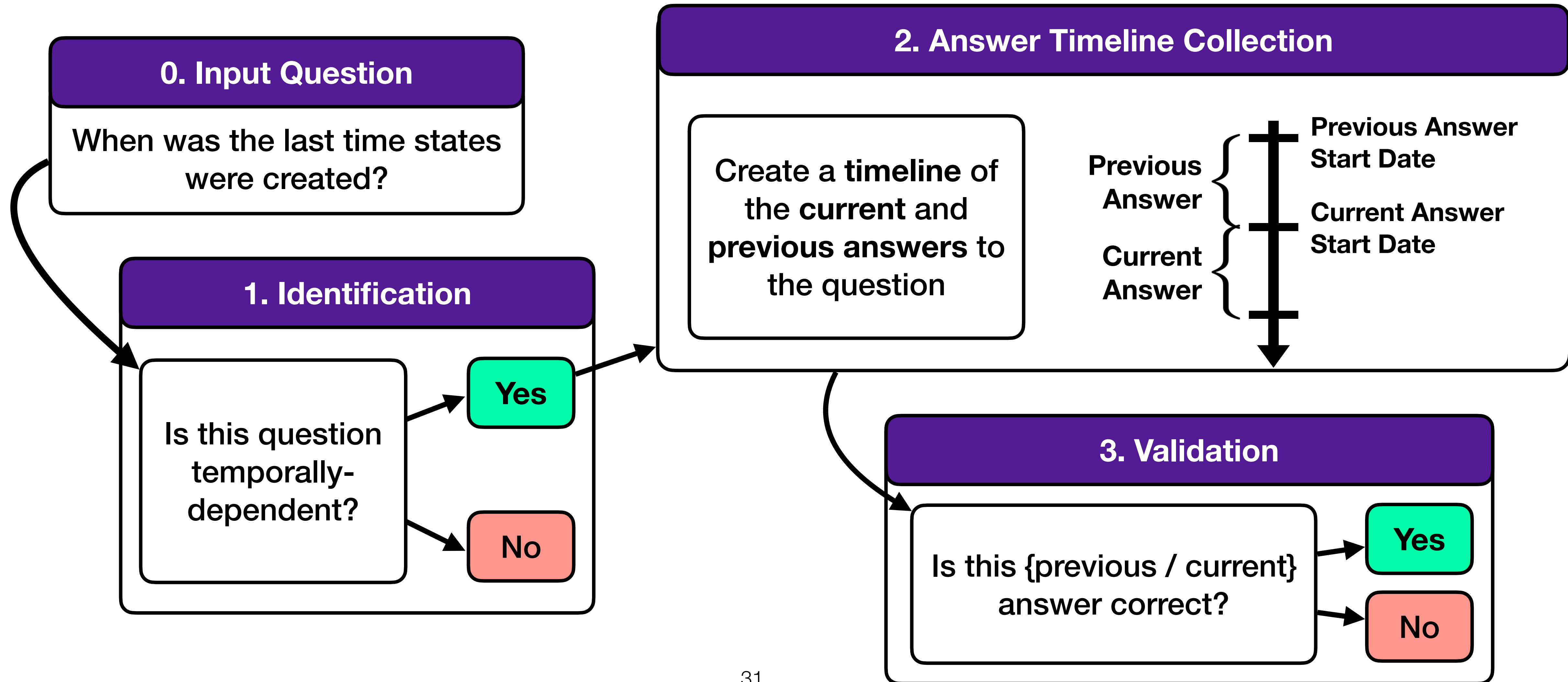
Dataset: Geographical SituatedQA



Dataset: Geographical SituatedQA

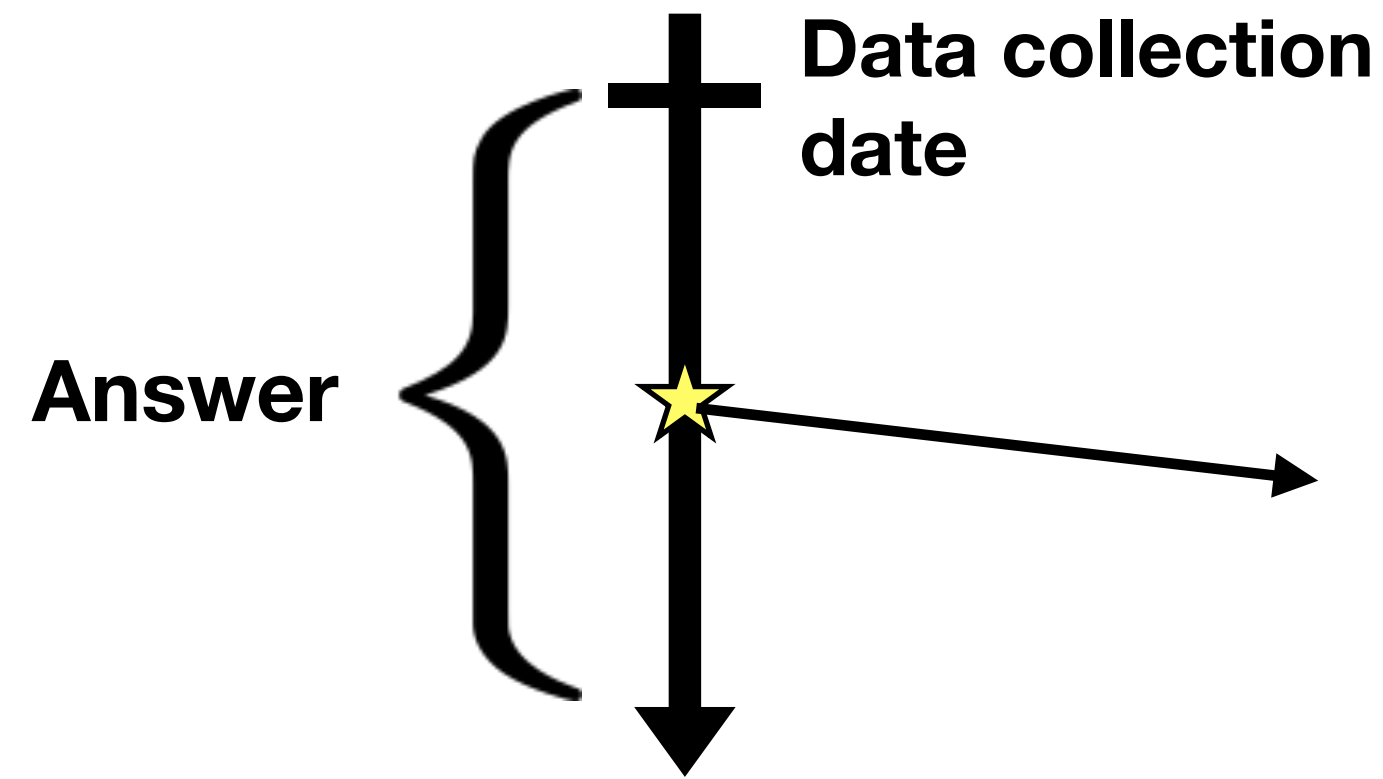


Dataset: Temporally SituatedQA



Mapping Answer Timeline to {Temporal Context / Answer}

Static Questions

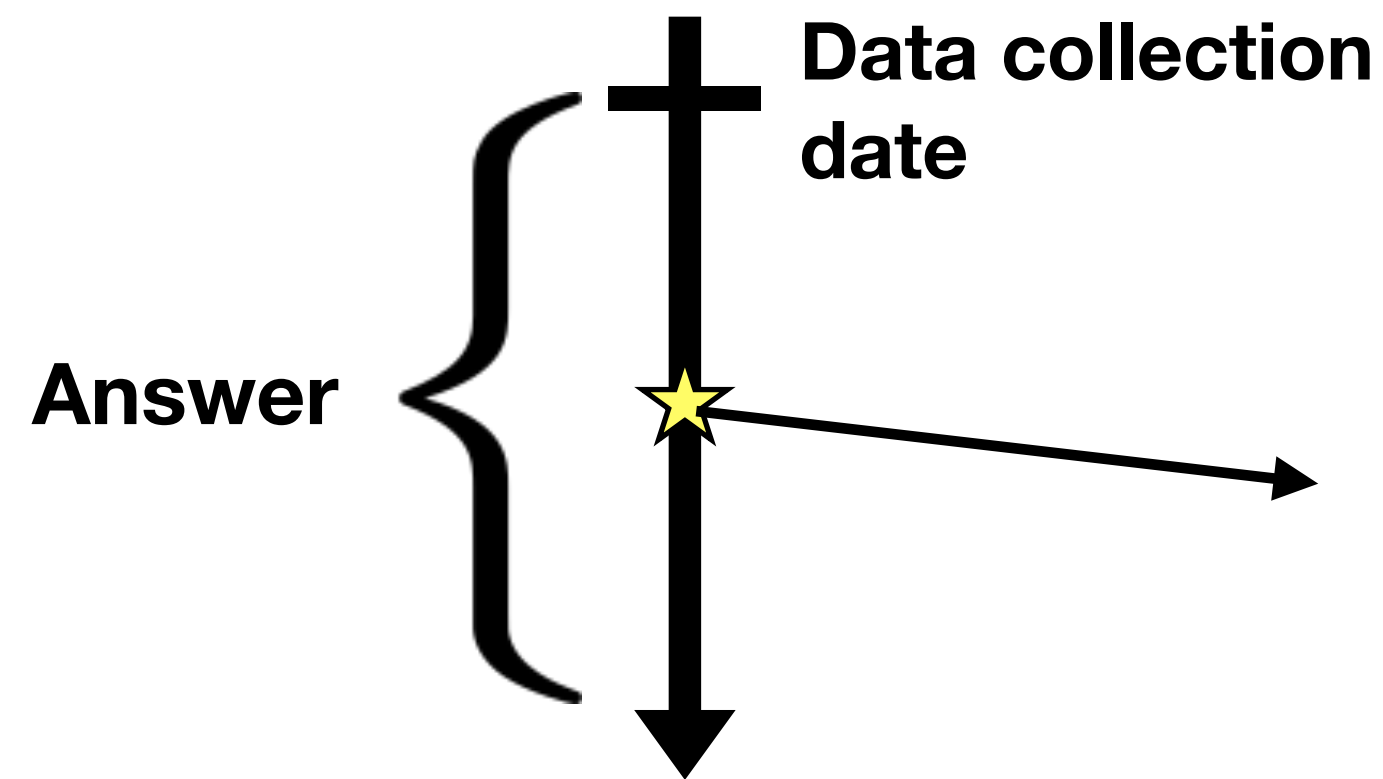


Static: Using temporally-independent questions by sampling a date after its original answer was collected

Example: Where is Belize as of **2020**?

Mapping Answer Timeline to {Temporal Context / Answer}

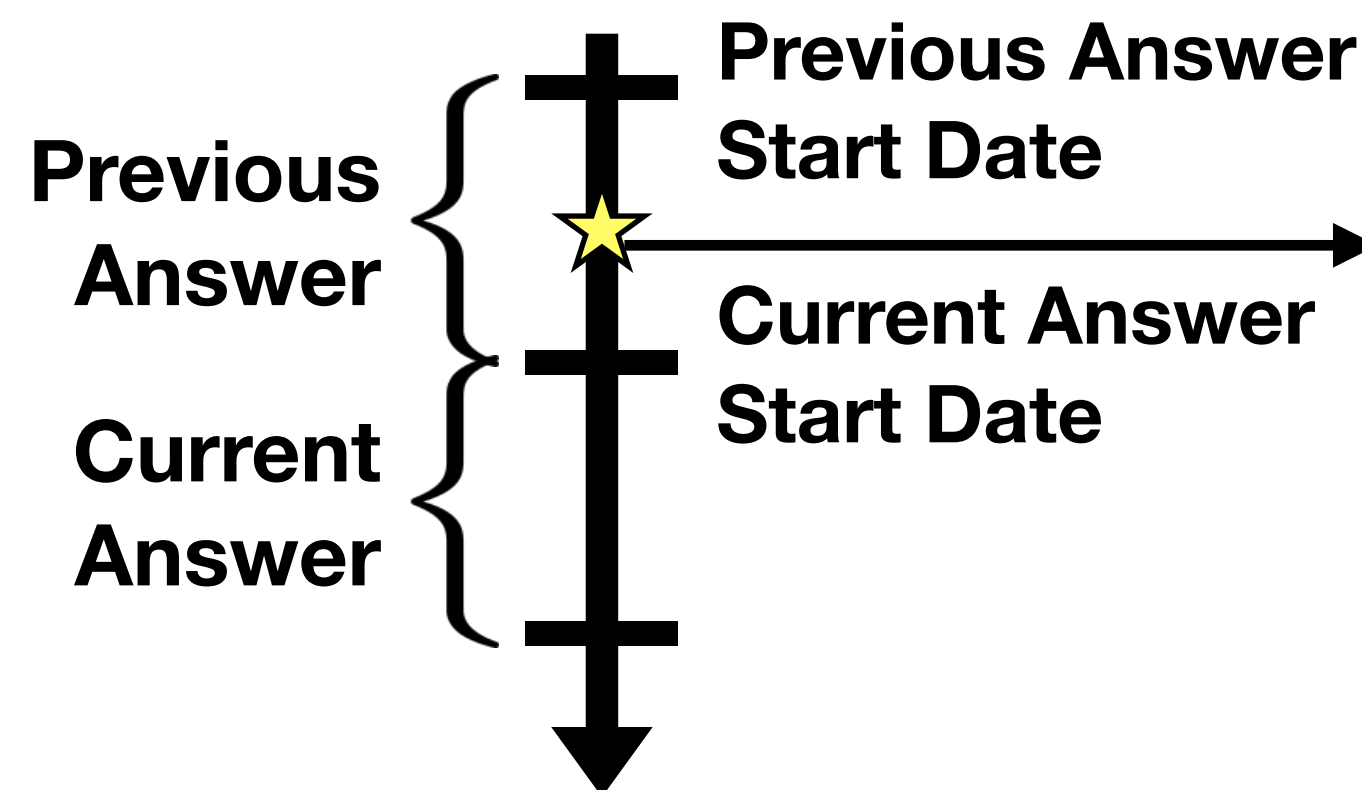
Static Questions



Static: Using temporally-independent questions by sampling a date after its original answer was collected

Example: Where is Belize as of **2020**?

Temporally Dependent Questions

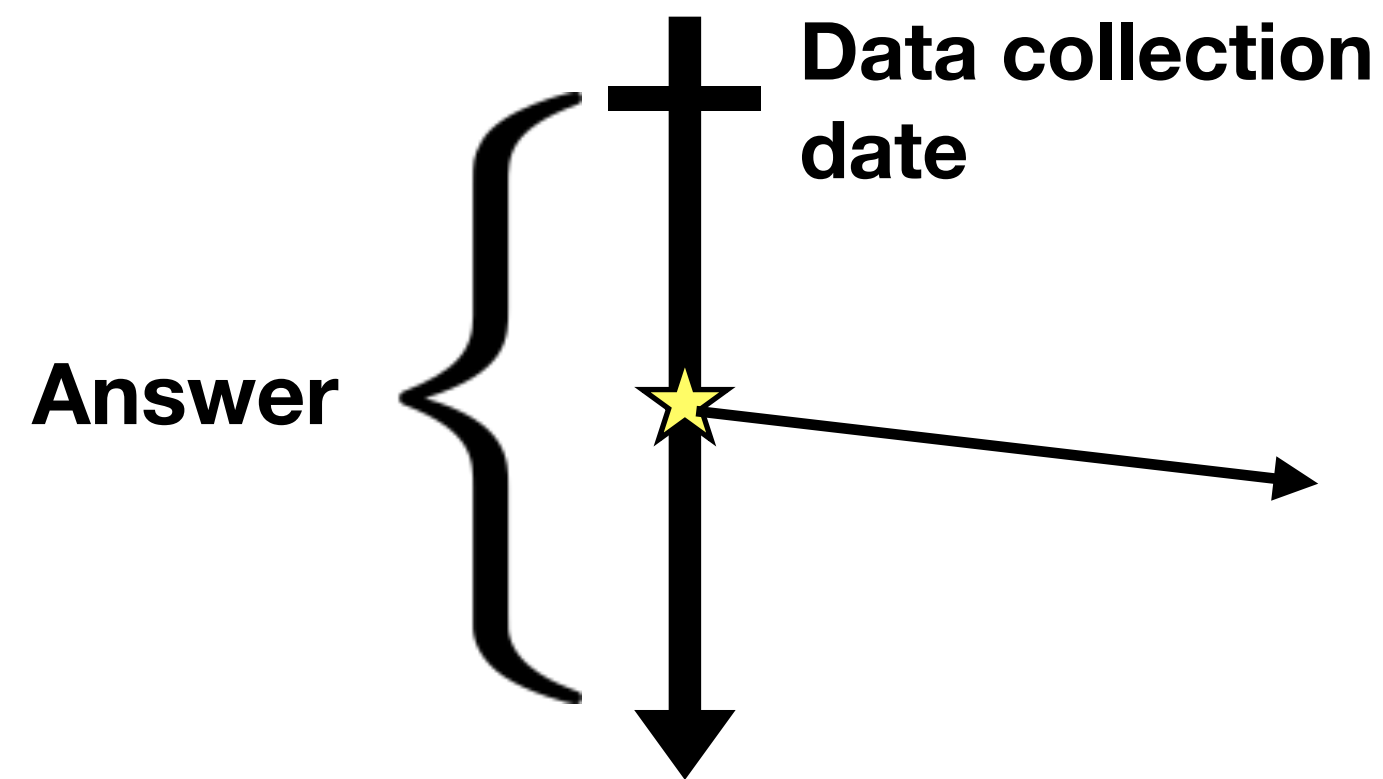


Sampled: selecting a timestamp between an answer's start and end date

Example: Who was the most recently appointed supreme court justice as of **March 25, 2019**?

Mapping Answer Timeline to {Temporal Context / Answer}

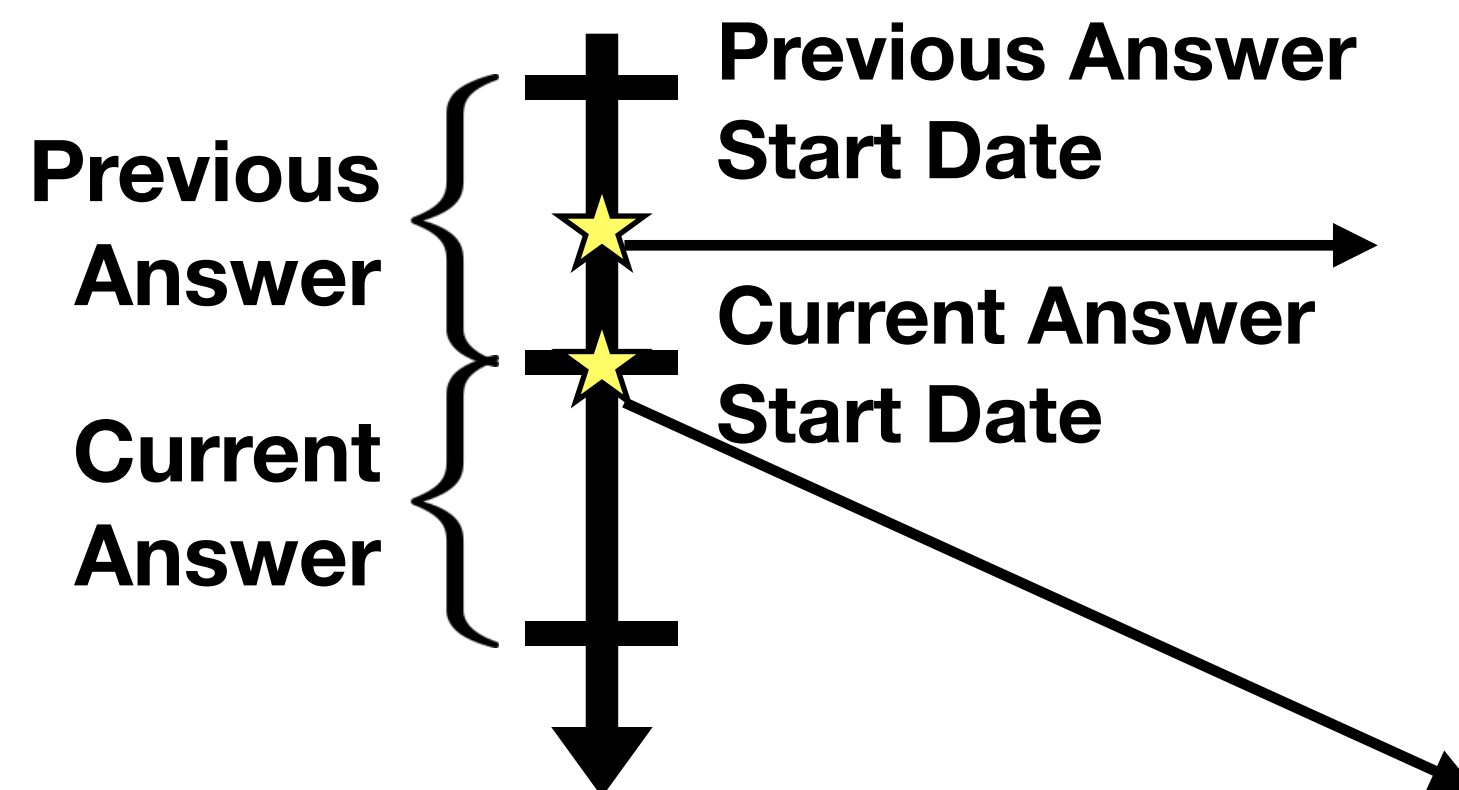
Static Questions



Static: Using temporally-independent questions by sampling a date after its original answer was collected

Example: Where is Belize as of **2020**?

Temporally Dependent Questions



Sampled: selecting a timestamp between an answer's start and end date

Example: Who was the most recently appointed supreme court justice as of **March 25, 2019**?

Start: using an answer's start date

Example: Who was the most recently appointed supreme court justice as of **October 26th 2020**?

The SituatedQA Task: (question, context) → answer

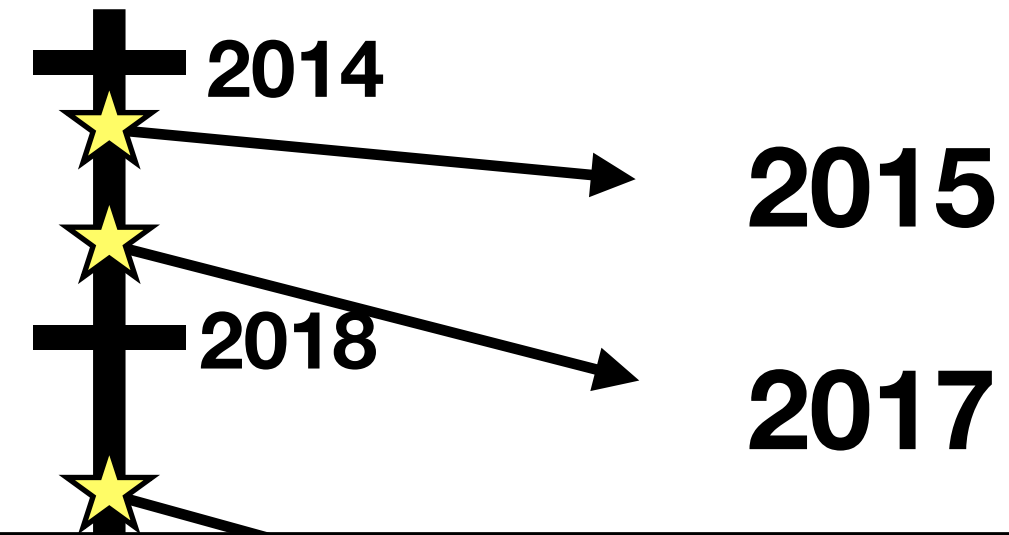


Question

Where was the last Winter Olympic Games held?



Temporal Context



Answer

Sochi

Sochi

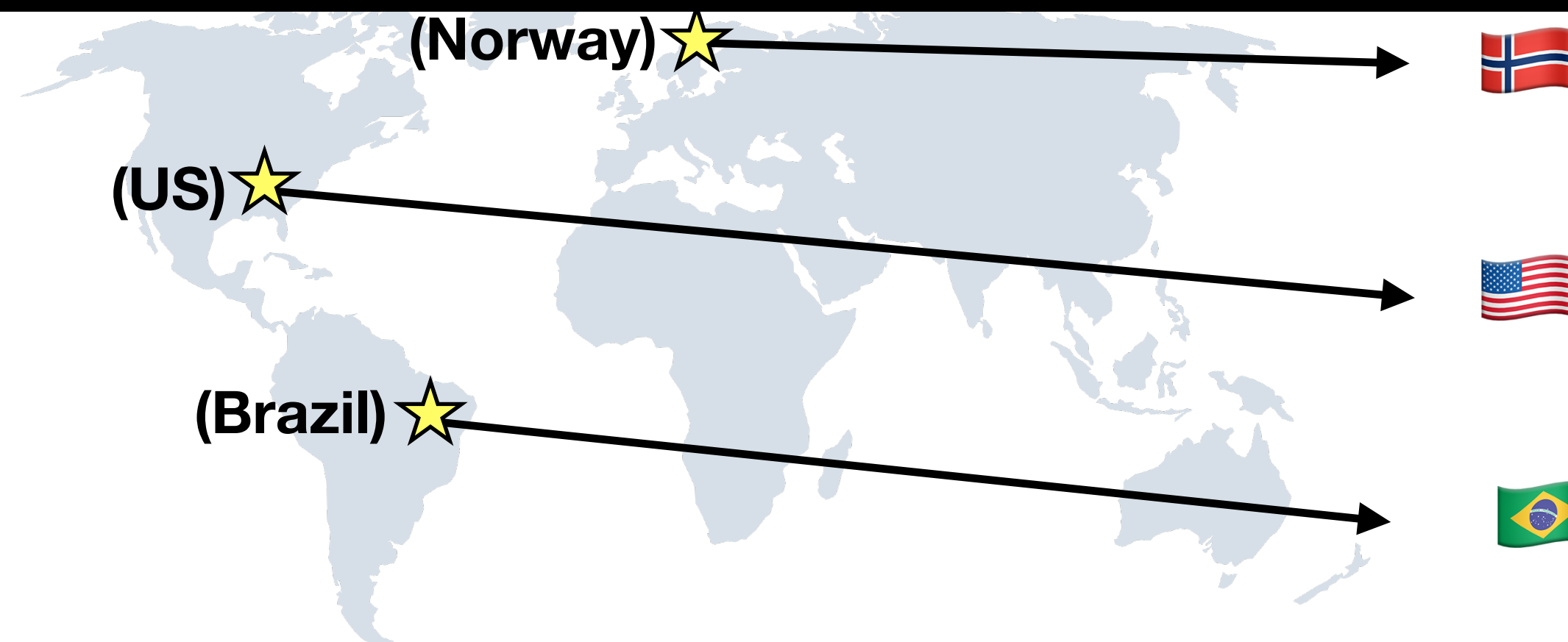
Chang

Sourcing from NQ-Open, we create:

- **Temporal SituatedQA** containing **6.7K examples**
- **Geographical SituatedQA** containing **5.5K examples**



How many gold medals did we win in the Winter Olympics 2018?



Answer

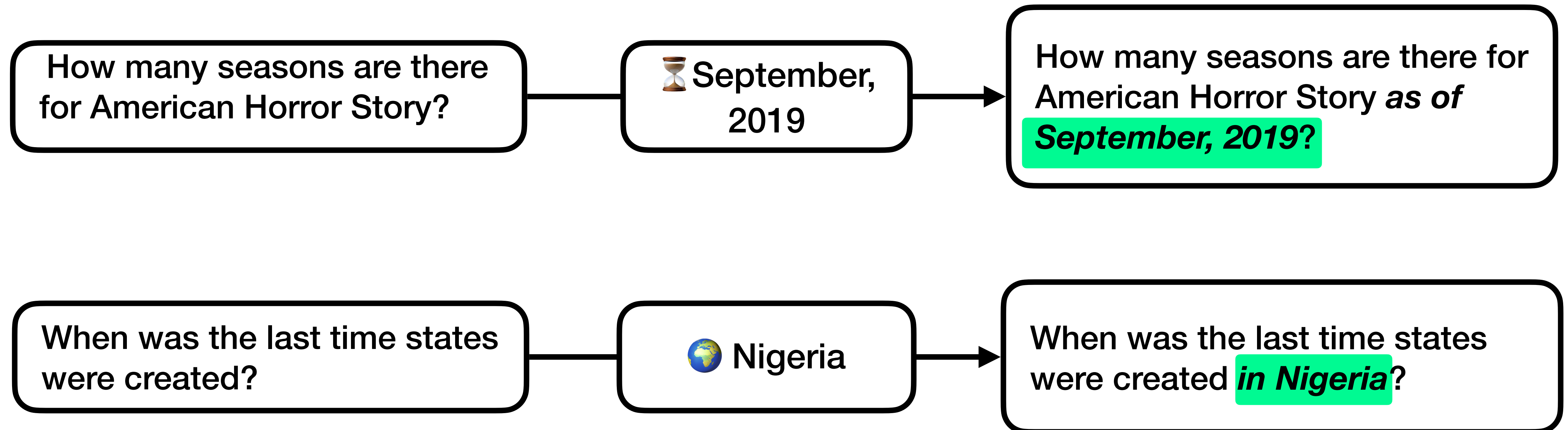
14

9

0

Experiments: Input modification

Query modification is used to specify extra-linguistic context to a model



Experiments: Models

Closed-Book (BART)



I remember 🧠!
The answer is _____

Retrieval-Based (DPR)



You found me 🕵️!
The answer is _____

Experiments: Models

Closed-Book (BART)



I remember 🧠!
The answer is _____

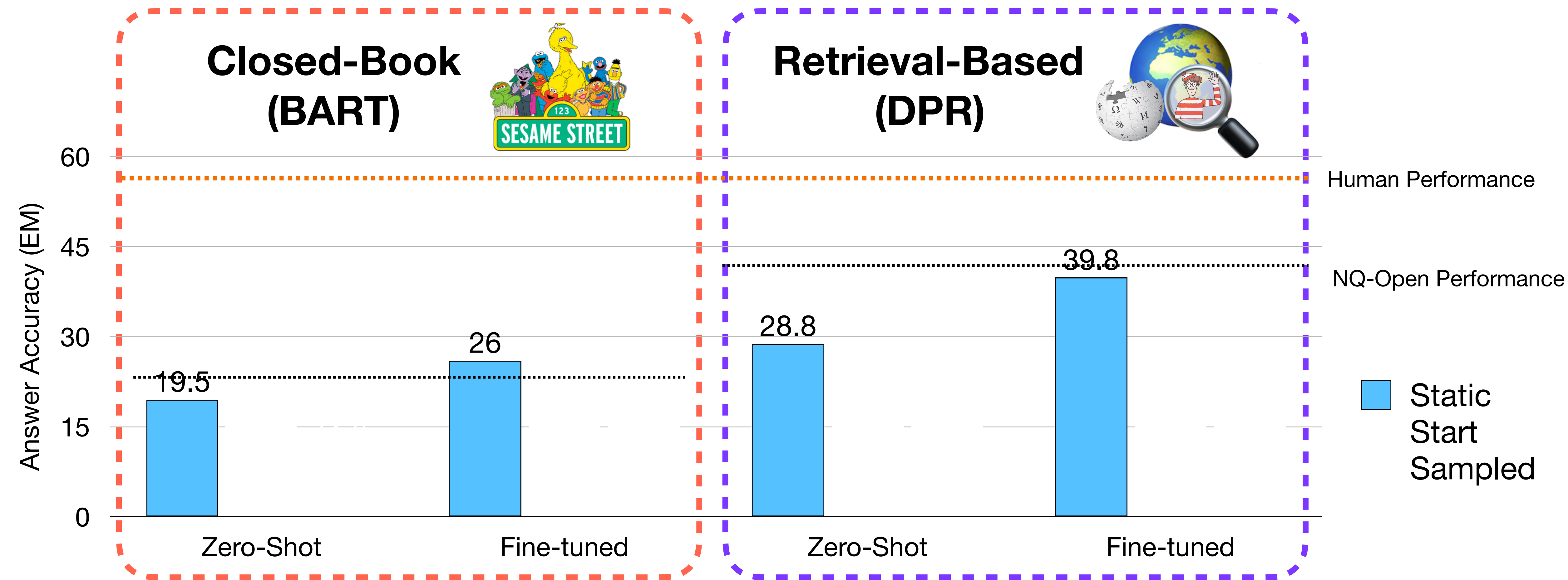
Retrieval-Based (DPR)



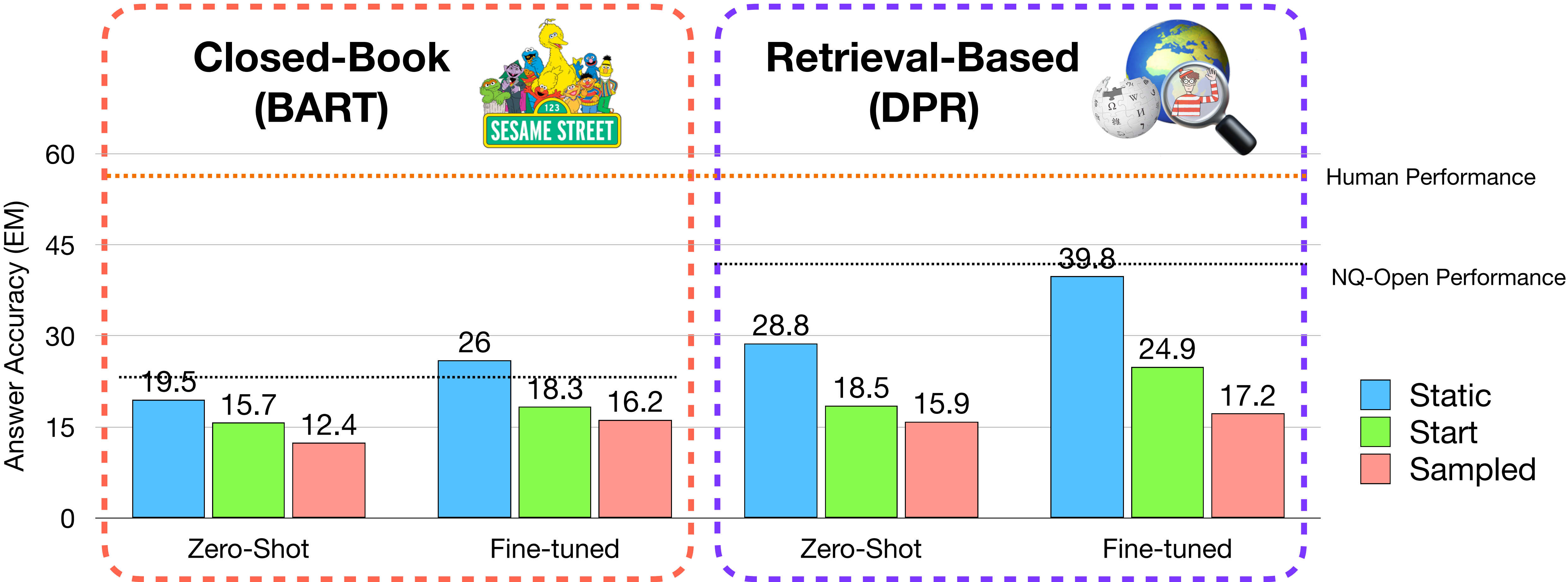
You found me 🕵️!
The answer is _____

- **Zero-shot** models are trained on NQ-Open
- **Fine-tuned** models are first trained on NQ-Open, then also tuned on our collected training data

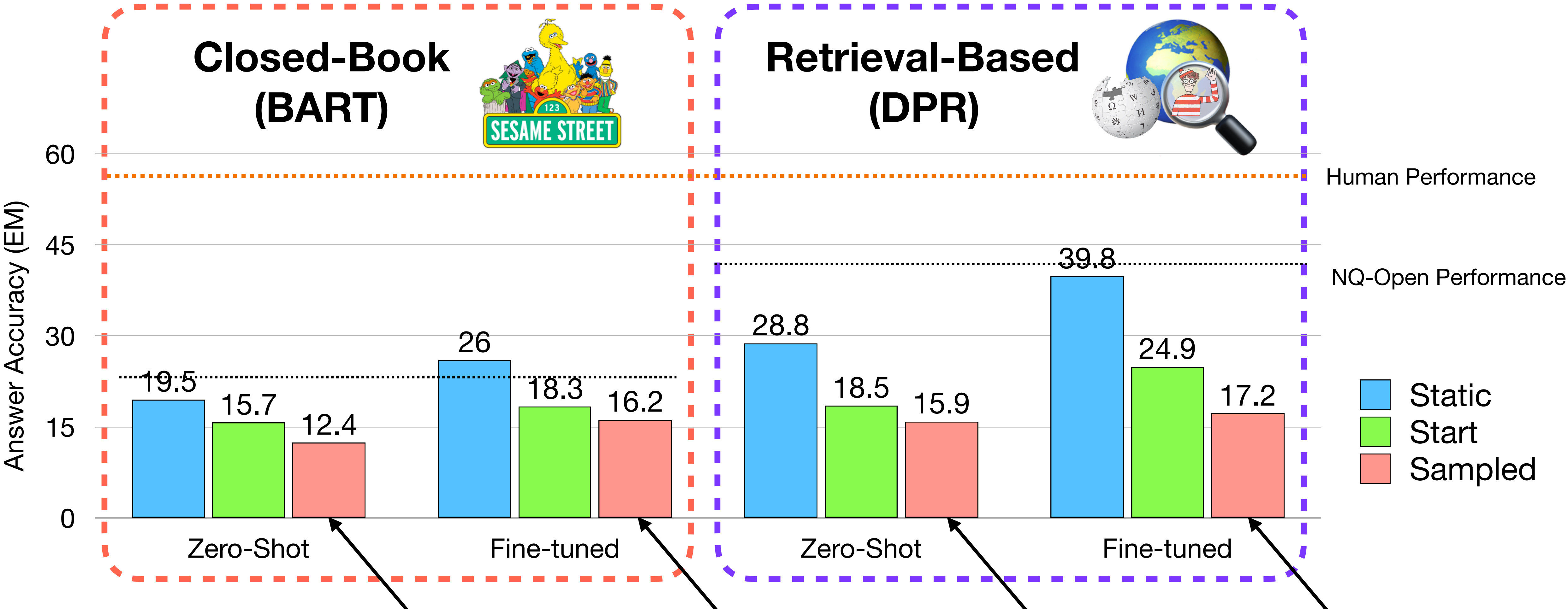
Results: Temporal SituatedQA



Results: Temporal SituatedQA



Results: Temporal SituatedQA



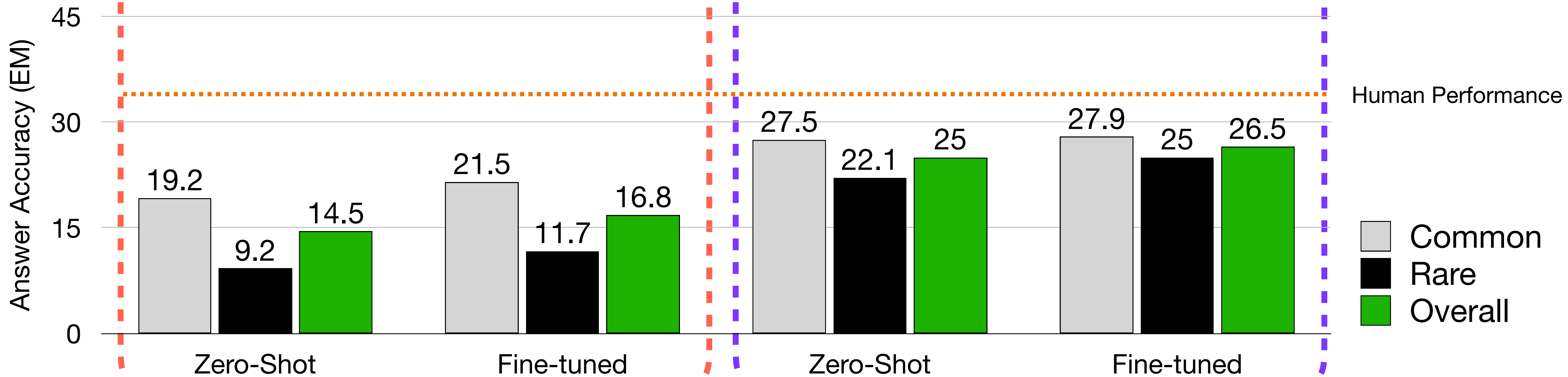
Models struggle the most with contexts that are *between* answer transitions

Results: Geographical SituatedQA

Closed-Book (BART)



Retrieval-Based (DPR)



Results: Geographical SituatedQA

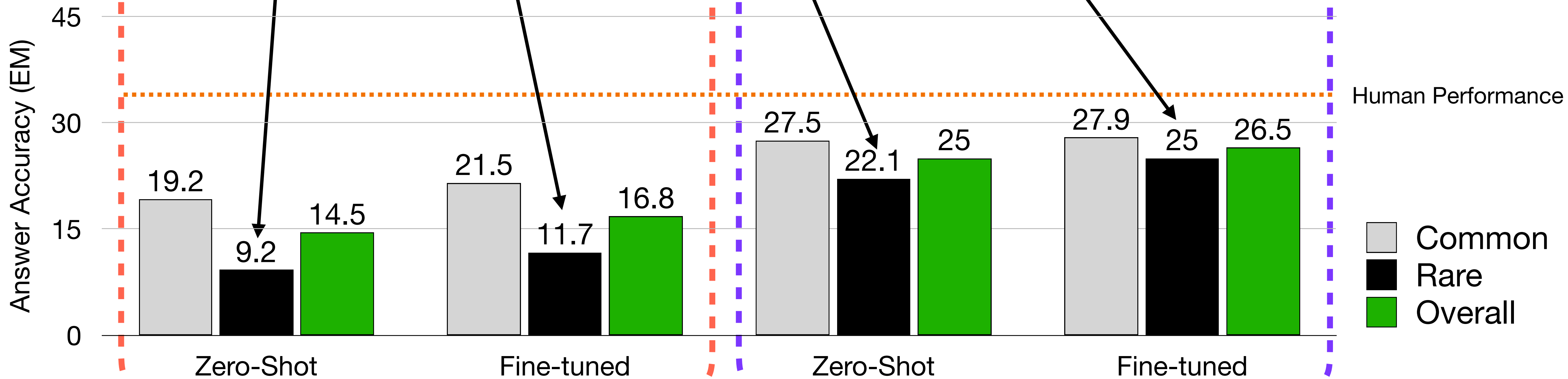
Closed-Book (BART)



Retrieval-Based (DPR)



1. Models perform worse on rare locations



Results: Geographical SituatedQA

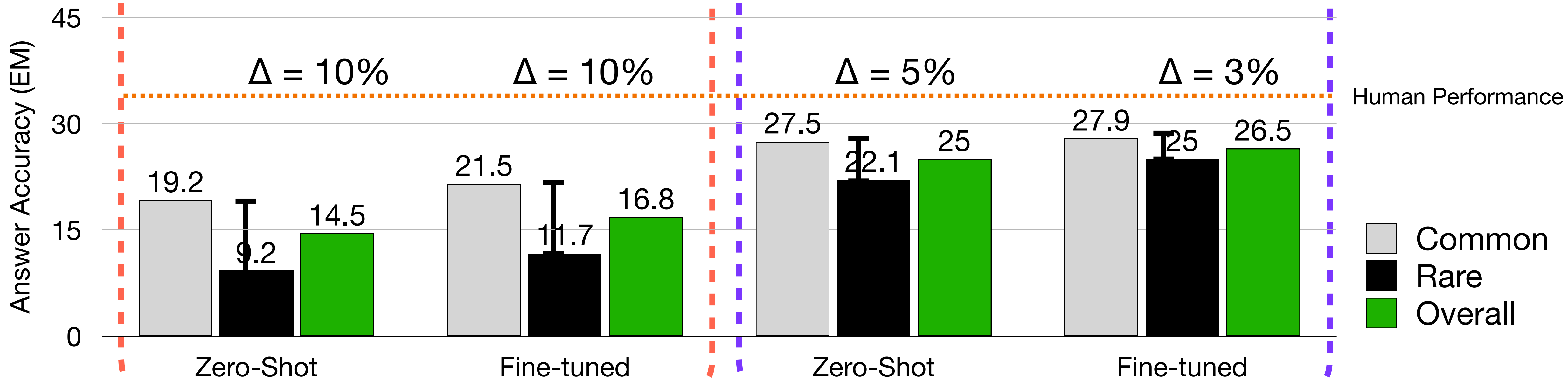
Closed-Book (BART)



Retrieval-Based (DPR)



1. Models perform worse on rare locations
2. The performance gap between rare and common locations is less significant for retrieval-based methods



Why is human performance so low?

- Automatic exact match score **underestimates** model's performance
- System showed up to 41–54% accuracy gain with human evaluation
 - **Alternative interpretation** of the question:
 - “who has the most superbowl rings”
 - “which person (including coaches) has the most superbowl rings”
 - “which player has the most superbowl rings”
 - “which team has the most superbowl rings”
 - Different **granularity** of answers:
 - the year (1982) vs. the month (October 1982)
 - the city (Pawtucket) vs. the state (Rhode Island).
 - Same answer, different **surface forms**:
 - “about 930 BCE” and “around 930 BC”

For more discussions, please look at our EfficientQA Workshop report! [Min et al, JMLR 2021]

Analysis from SituatedQA data



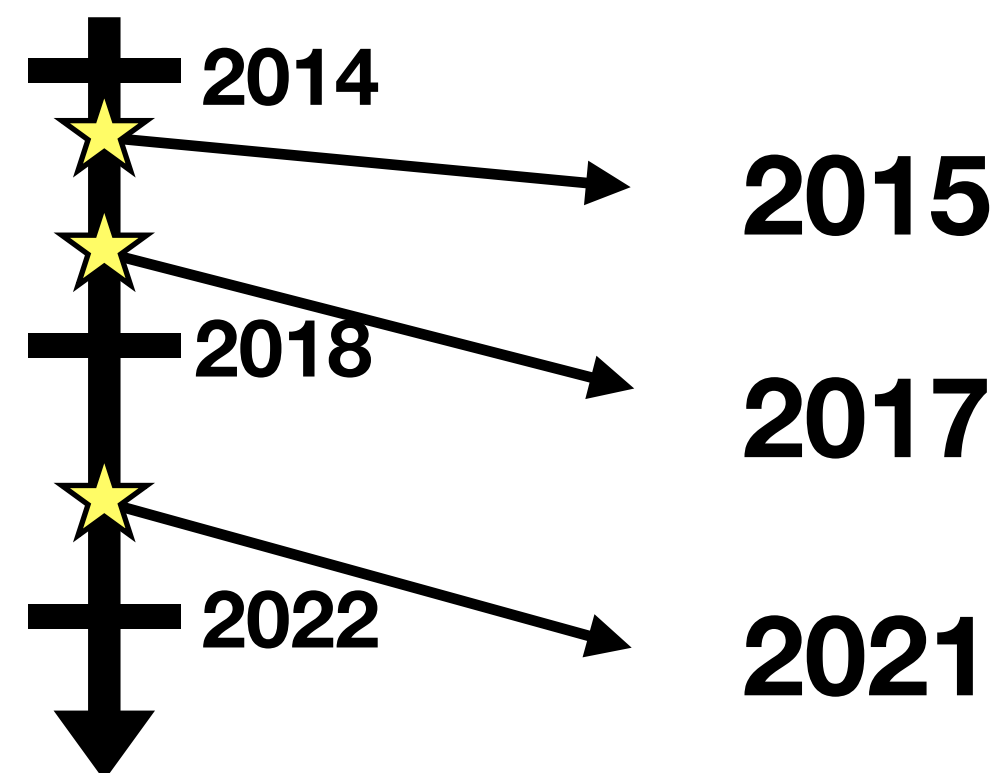
Question

Where was the last Winter Olympic Games held?



,

Temporal Context



Answer

Sochi

Sochi

Pyeongchang



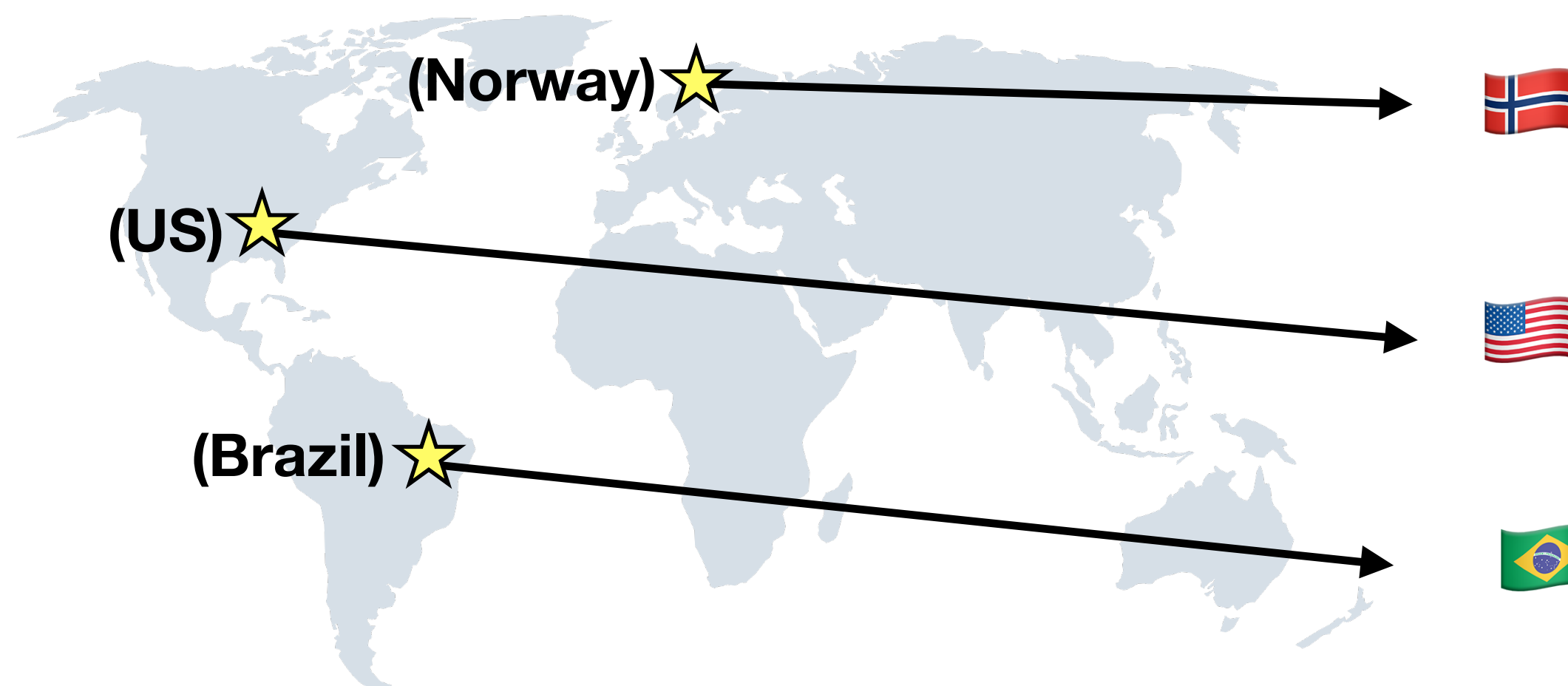
Question

How many gold medals did we win in the Winter Olympics 2018?



,

Geographical Context



Answer

14

9

0

Analysis: Geographical Bias

Q: What is the assumed context for a geographically-dependent question?

- Provide model with purposefully ambiguous question without geographical location specified, which location's answer would the model output?

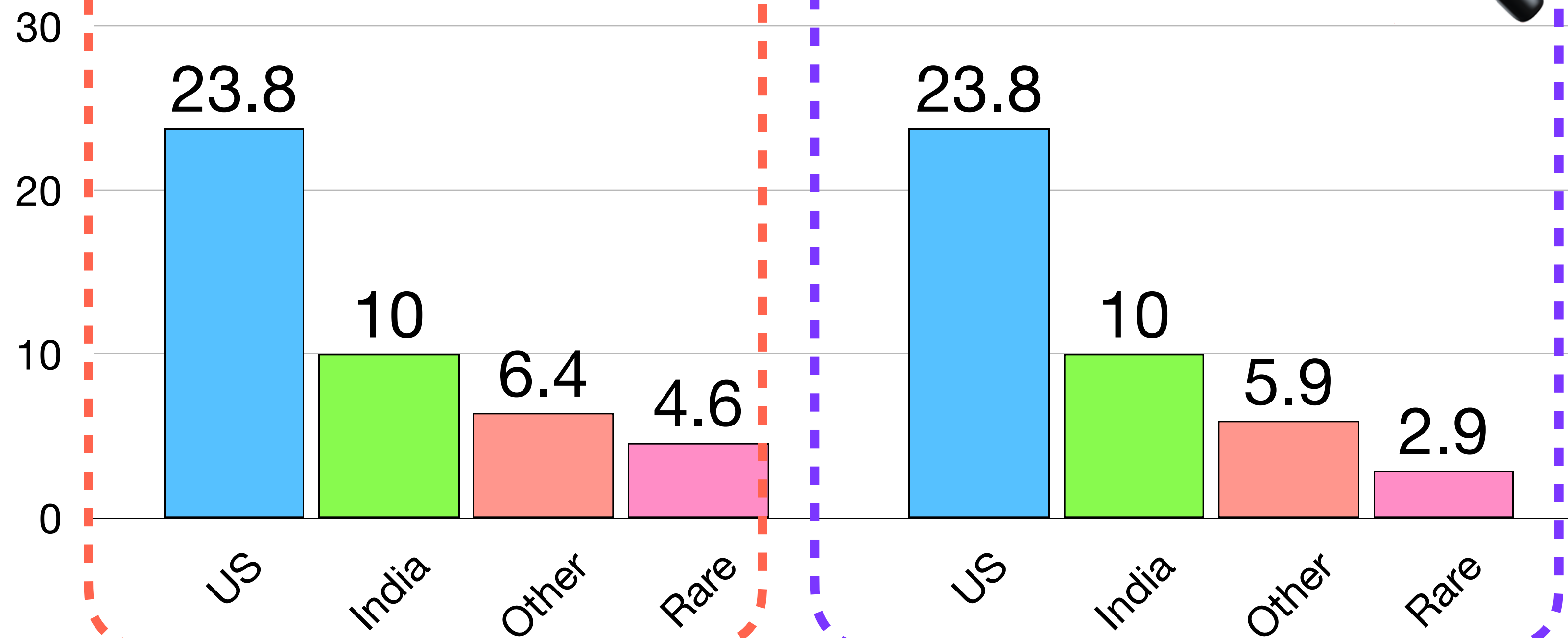
**Closed-Book
(BART)**



**Retrieval-Based
(DPR)**



Accuracy (Exact Match)



Analysis: Geographical Bias

Q: What is the assumed context for a geographically-dependent question?

- Provide model with purposefully ambiguous question without geographical location specified, which location's answer would the model output?

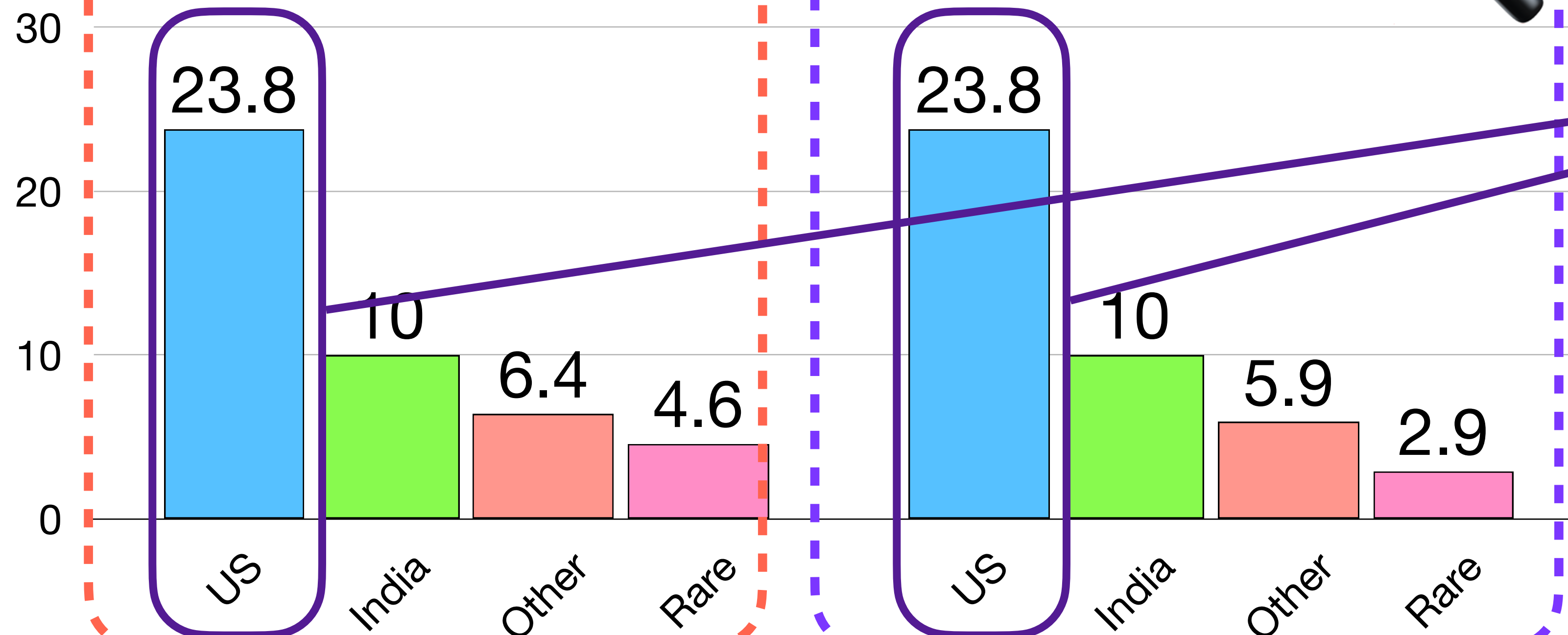
**Closed-Book
(BART)**



**Retrieval-Based
(DPR)**



Accuracy (Exact Match)

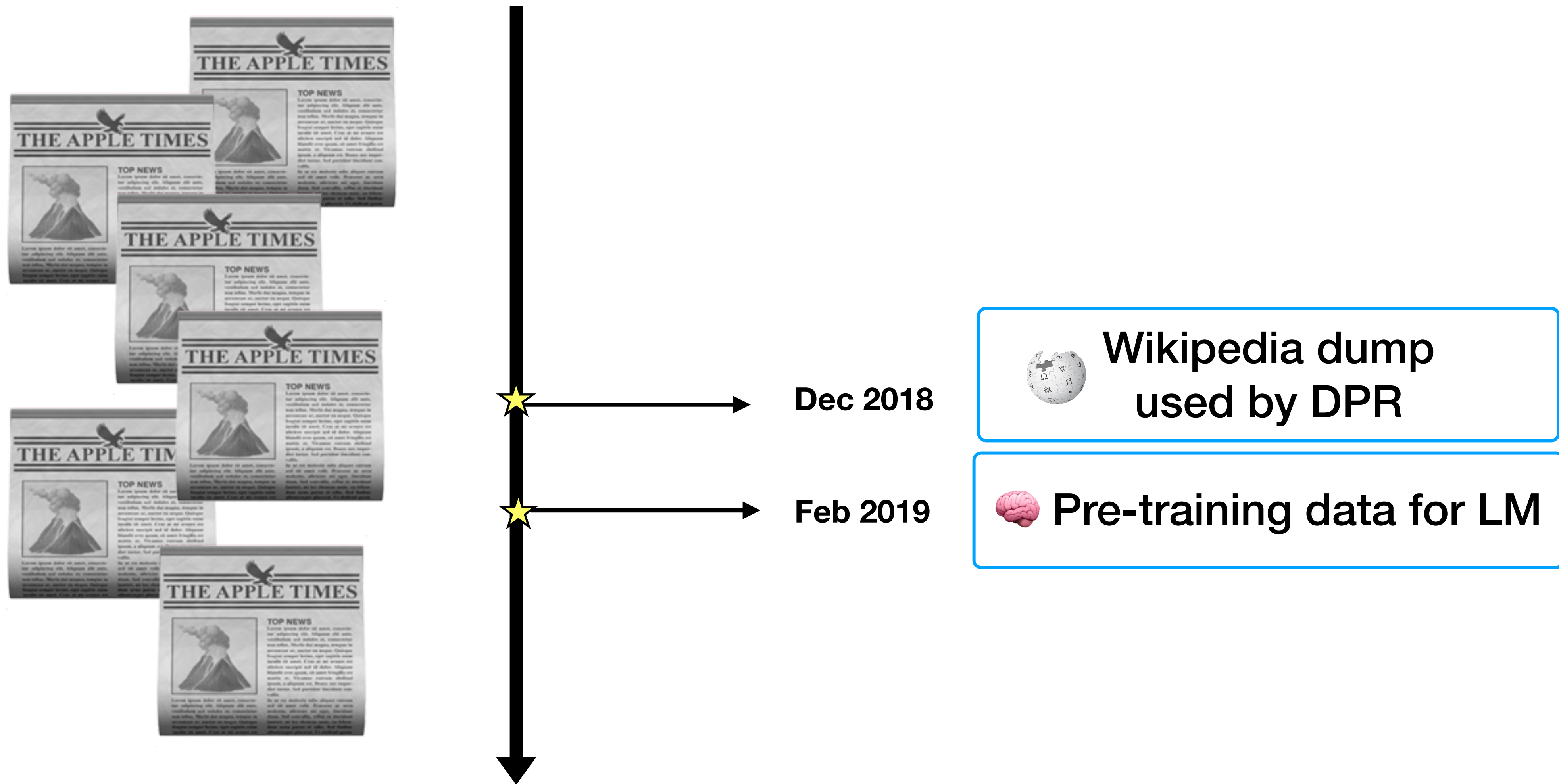


Models tend to assume the question is posed in the US

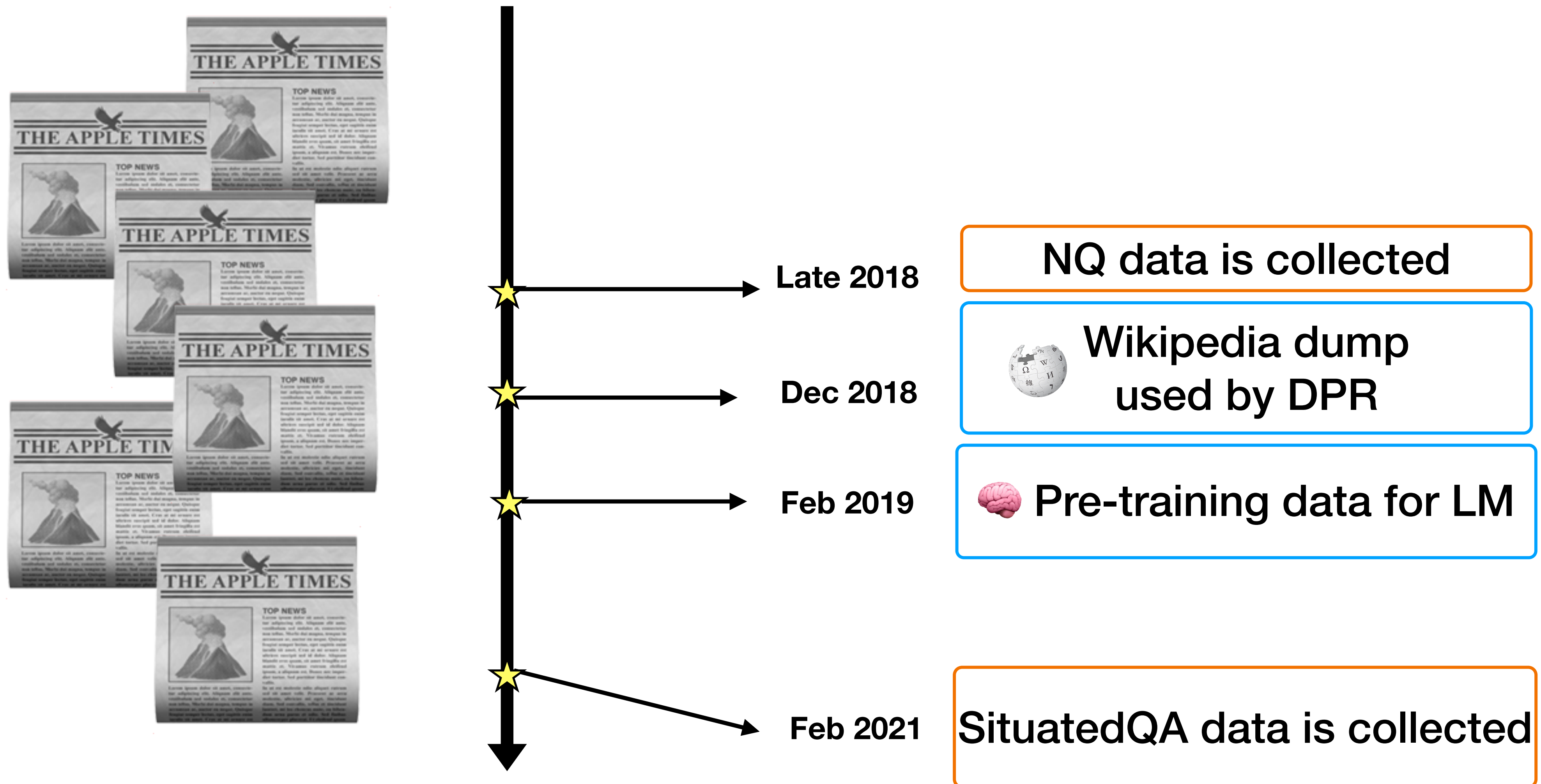
Analysis: Temporal Adaptation

Q: In original open retrieval QA setting, do models trained on data collected in the past generalize to answering questions asked in the present?

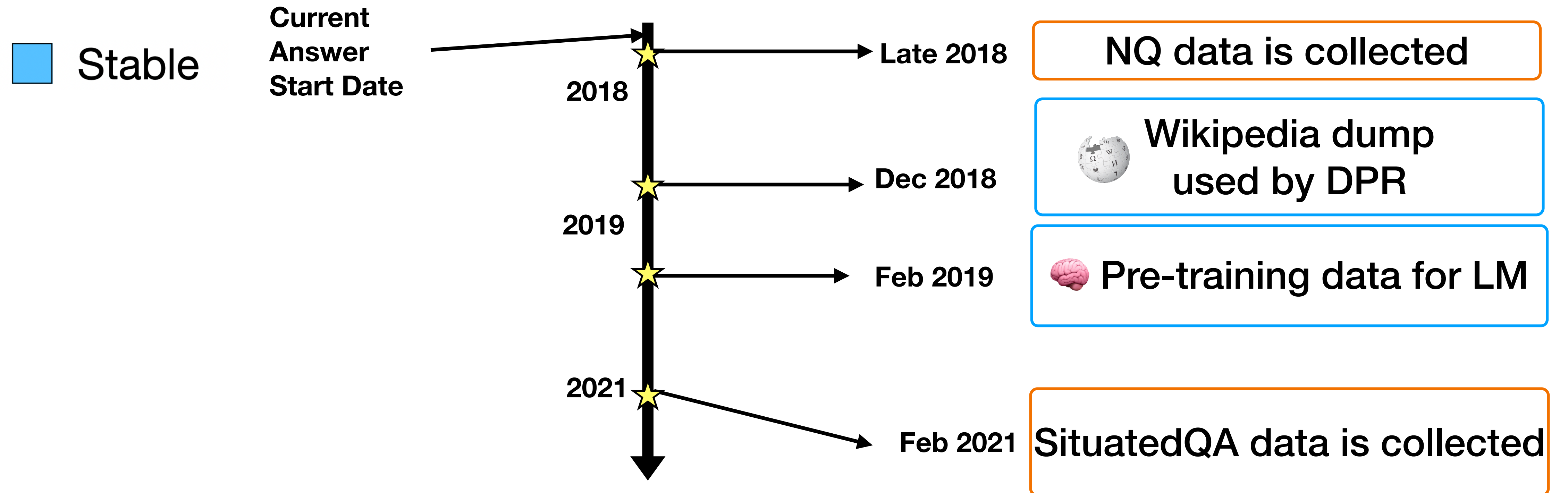
Temporal Dependence of Models



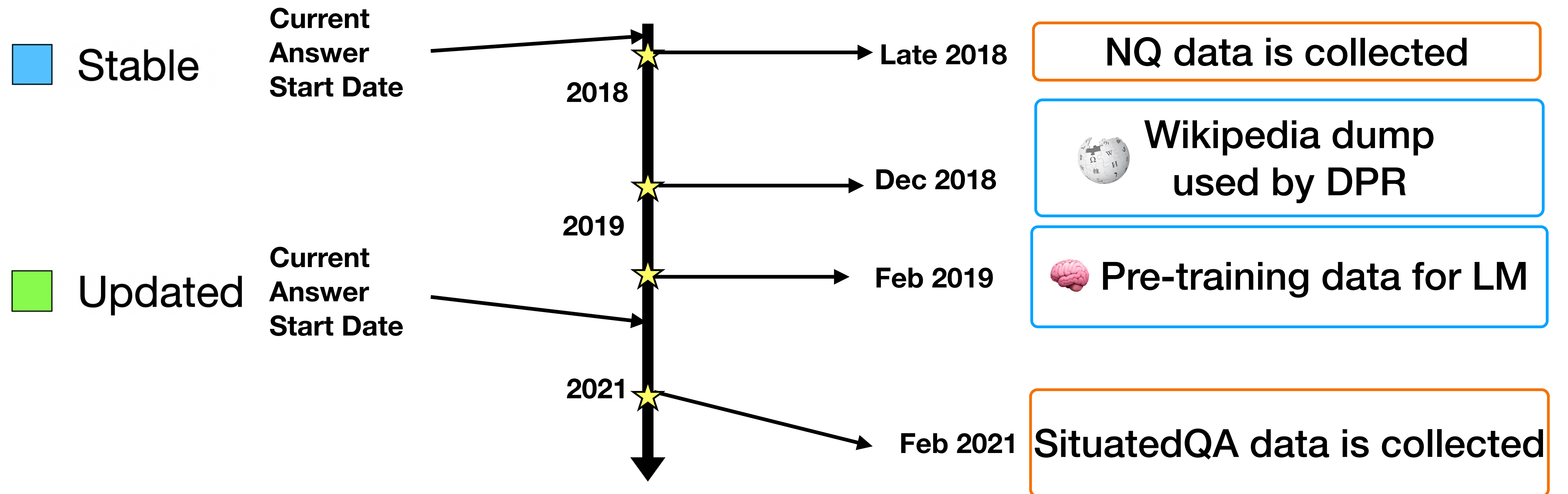
Temporal Dependence of Models



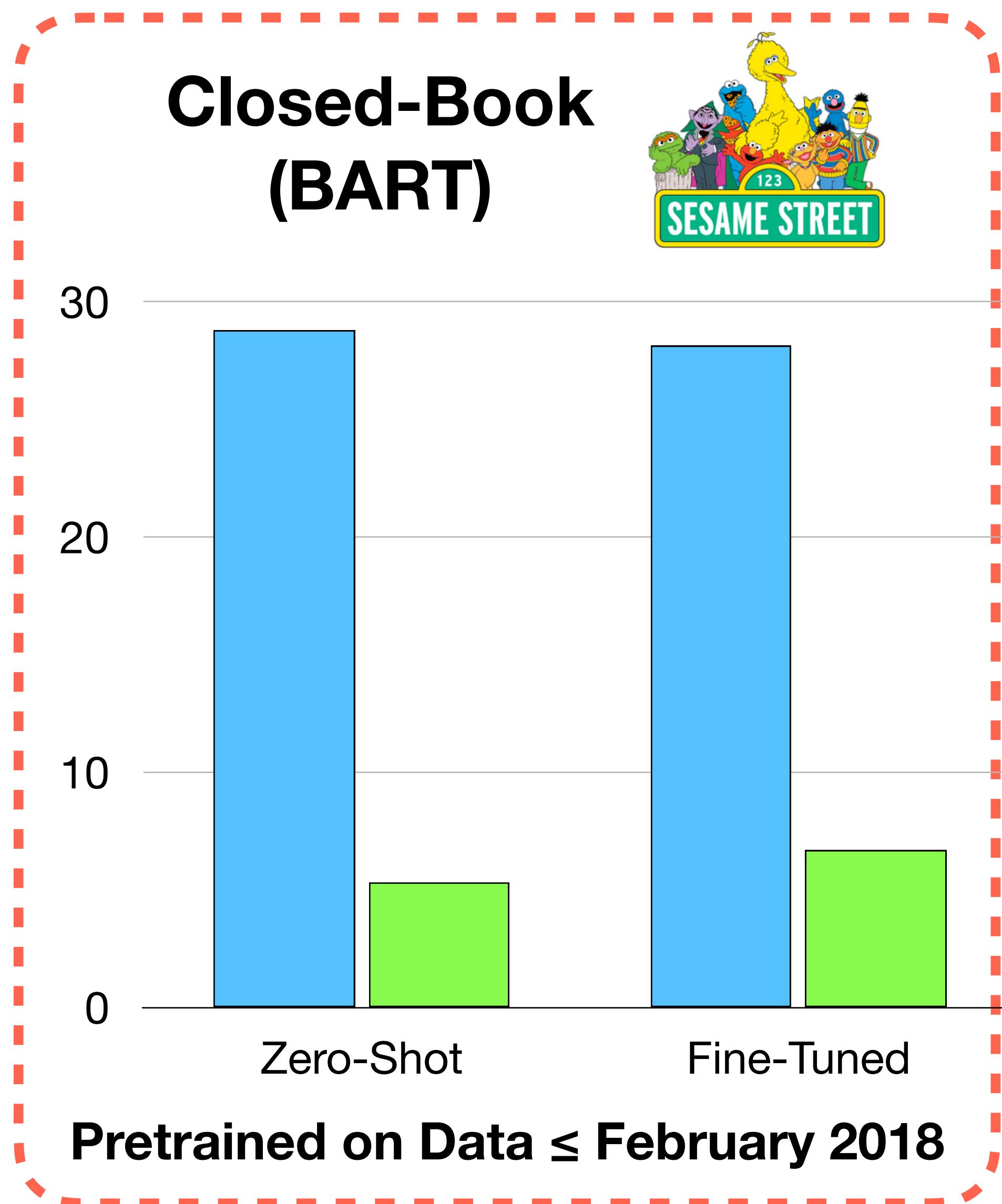
Analysis: Temporal Adaptation



Analysis: Temporal Adaptation



Analysis: Temporal Adaptation

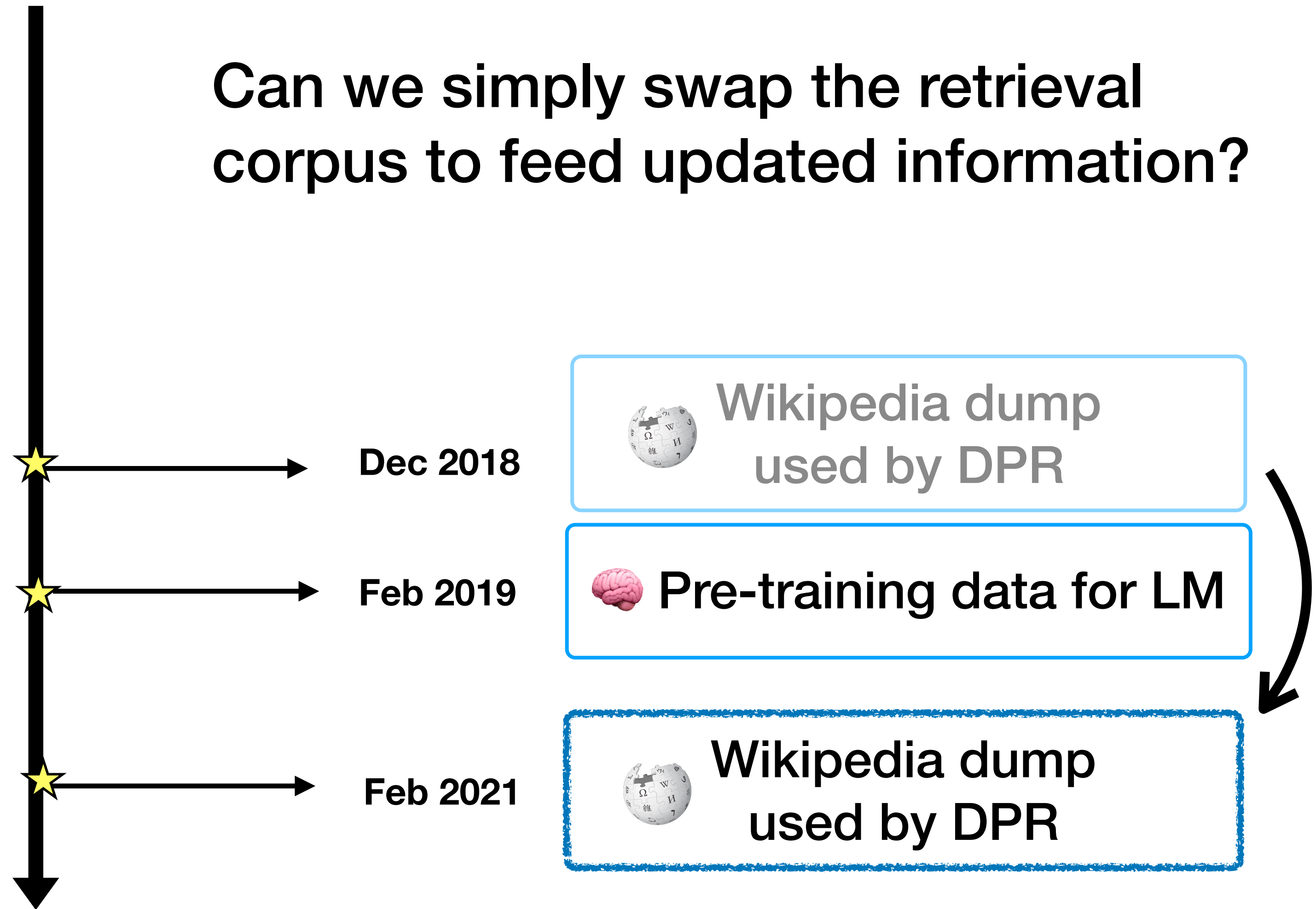


Closed-book models *cannot* adapt to questions with updated answers

Temporal Dependence of Models



Can we simply swap the retrieval corpus to feed updated information?

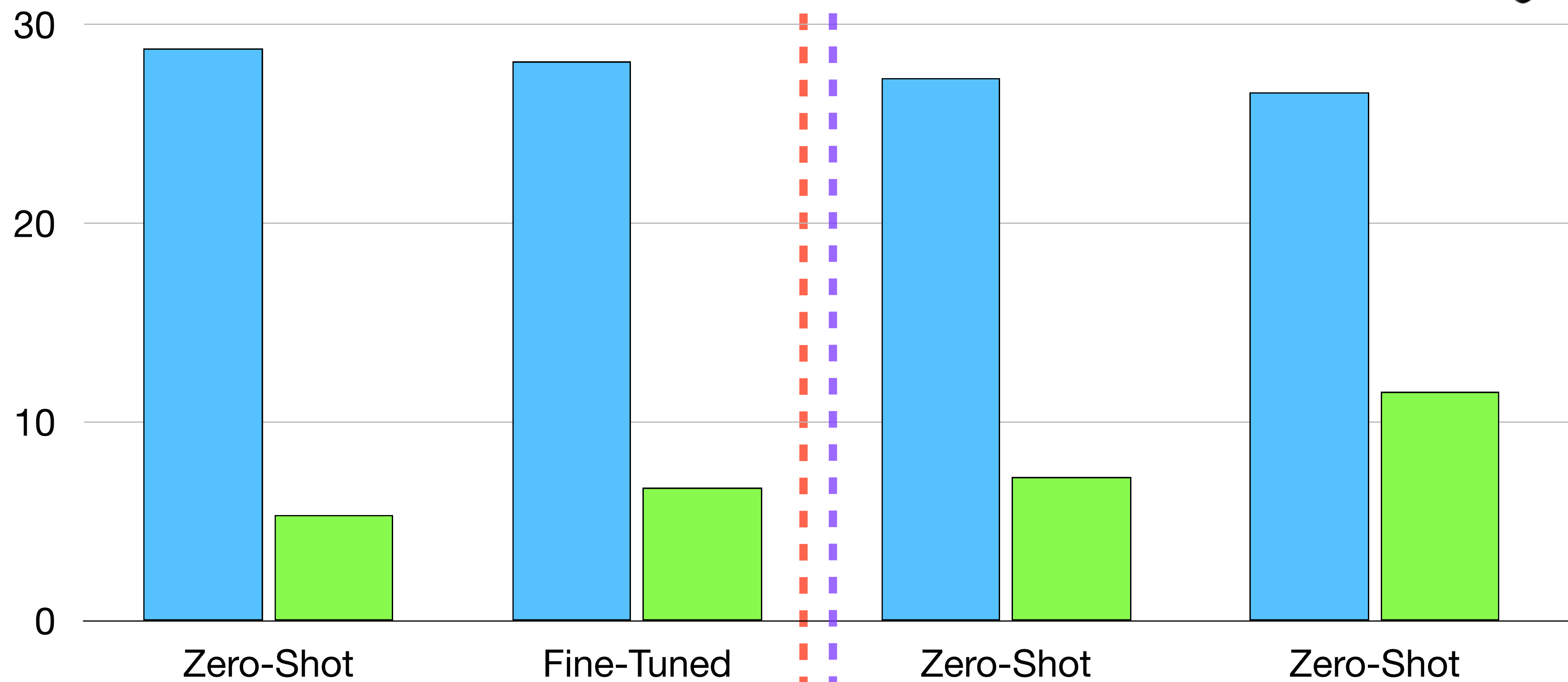


Analysis: Temporal Adaptation

Closed-Book (BART)



Retrieval-Based (DPR)



Pretrained on Data \leq February 2018

December 2018

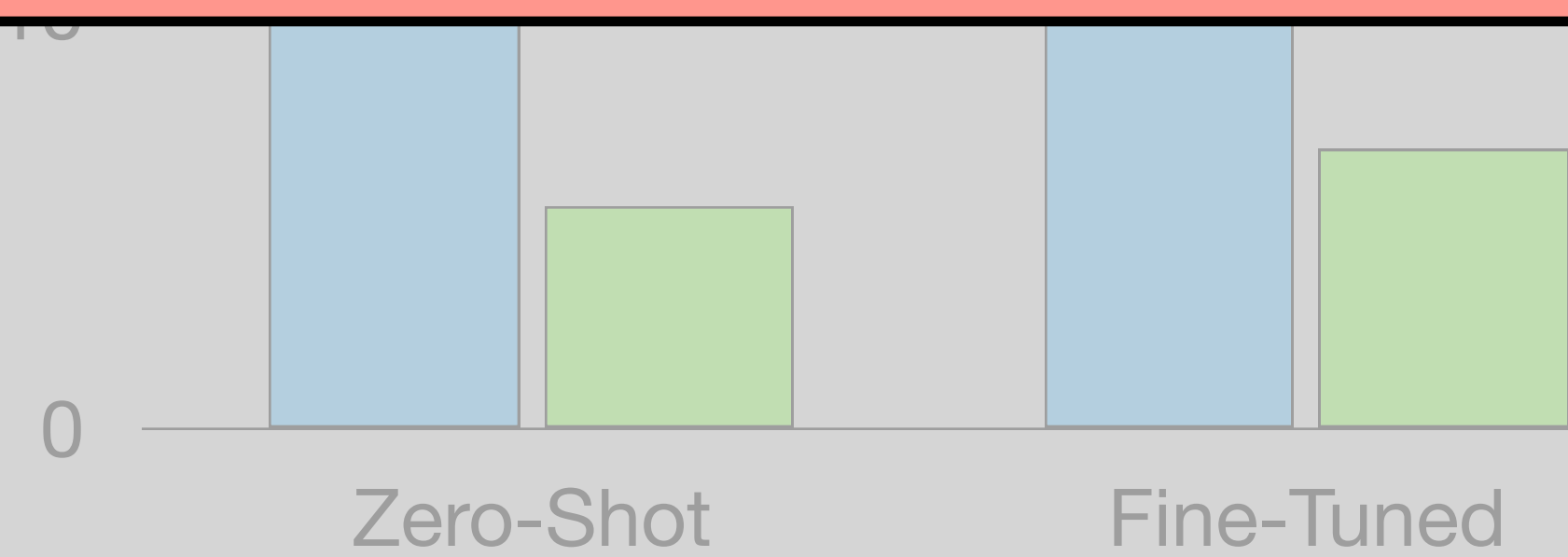
February 2021

Analysis: Temporal Adaptation

Closed-Book
(BART)



Retrieval-based methods
also fail to generalize,
*even with access to an
updated corpora*

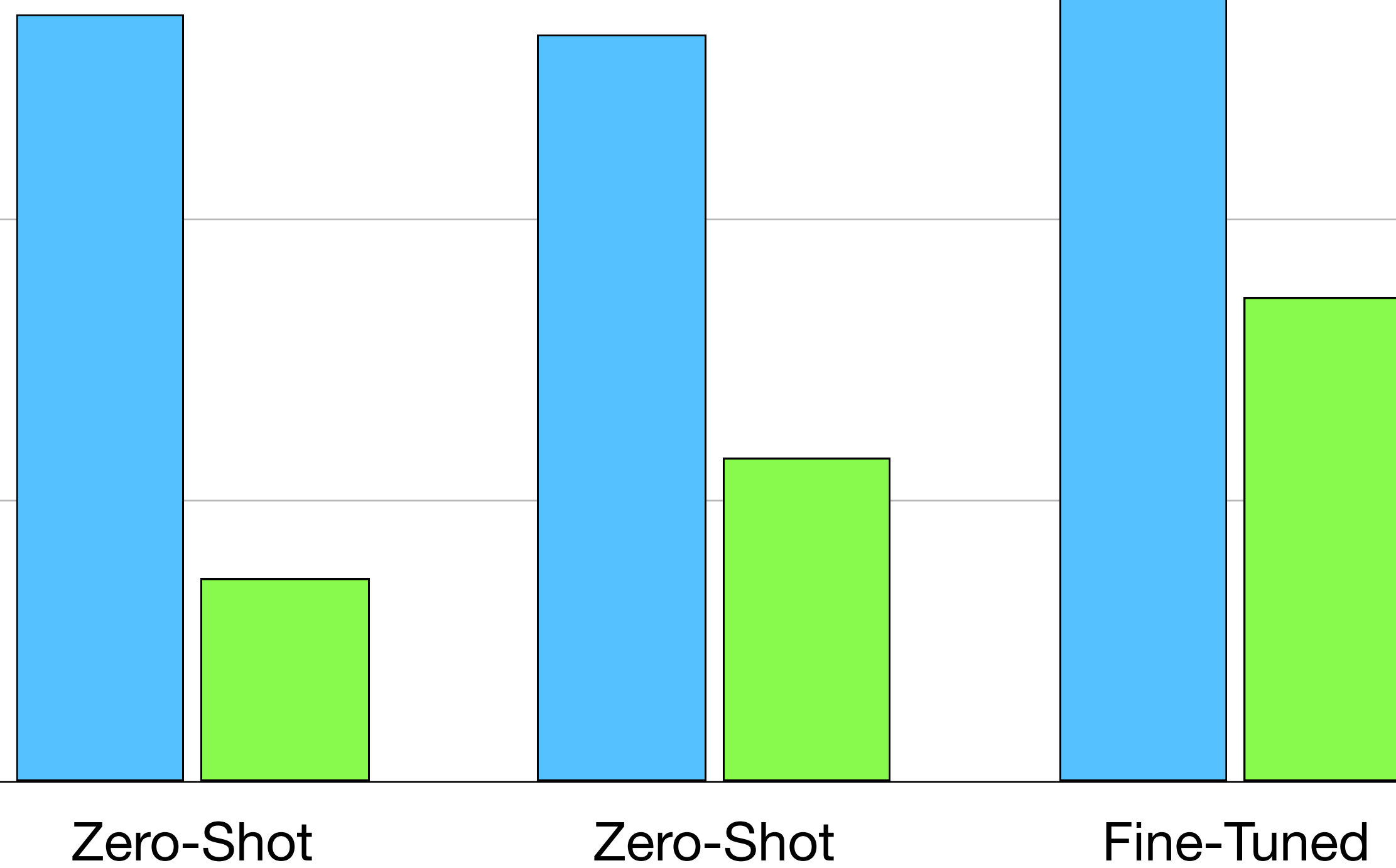


Pretrained on Data \leq February 2018

Retrieval-Based
(DPR)



Stable
Updated



December 2018

February 2021

Keeping Benchmark Temporally Valid

We can fix the world knowledge that we will query from:



[Petroni et al, NAACL 2021]

We will keep updating the test set with newly acquired examples:

The logo for TuringAdvice, consisting of the text "TuringAdvice" in white on a dark grey rectangular background.

TuringAdvice

[Zeller et al, NAACL 2021]



[Kiela et al, ArXiv 2021]

We can embrace temporal context into the task definition!

Related Work:

AmbigQA [Min et al, EMNLP 2020] explores ambiguity in QA, including ambiguity that cannot be resolved with context

[Lazaridou et al, ArXiv 2021, Dhingra et al, ArXiv 2021] explore creating temporally-aware pretrained language models

GD-VCR [Yin et al, EMNLP 2021] looks at geographical dependence in visual commonsense reasoning

This Talk

- Incorporating two extra-linguistic contexts (🕒 temporal and 🌍 geographical) into open retrieval question answering
- Presenting a benchmark which evaluates model's reasoning ability anchored at entity information

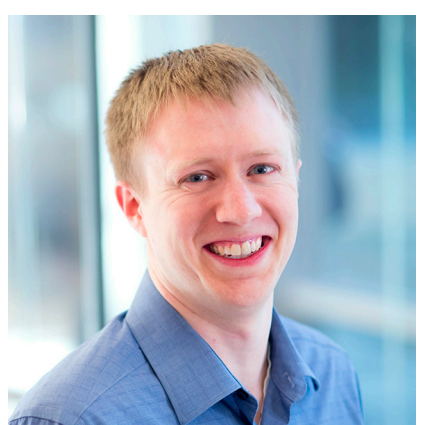
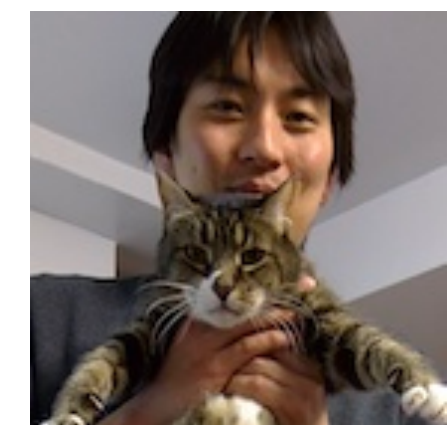
SituatedQA: Incorporating Extra-Linguistic Contexts into QA [EMNLP 2021]

Michael J.Q. Zhang and Eunsol Choi



CREAK: A Dataset for Commonsense Reasoning over Entity Knowledge [In submission]

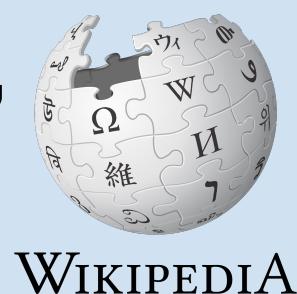
Yasumasa Onoe, Michael JQ Zhang, Eunsol Choi, Greg Durrett



Our work: reasoning based on facts about entities

Fact Verification [Vlachos and Riedel, 2014, Thorne et al., 2018]

- Simple facts on real world entities



Claim: There exists a producer and an actor called Simon Pegg

Supported

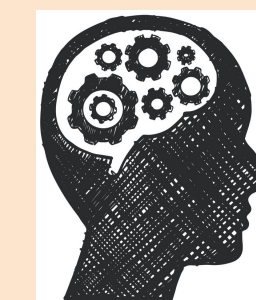
Claim: Mary McGee was the first woman to compete in the Baja 1000, between 1971 and 1979.

Refuted

Commonsense Reasoning

[Levesque et al., 2011, Talmor et al., 2019, 2021]

- Reasoning on fictional / general entities



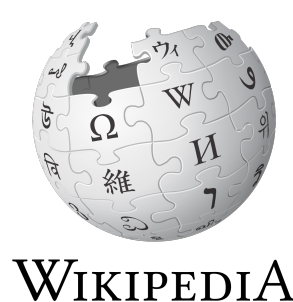
*Where on a **river** can you hold a cup upright to catch water on a sunny day?*

✓ **waterfall**, ✗ **bridge**, ✗ **valley**, ✗ **pebble**, ✗ **mountain**

*Where can I stand on a **river** to see water falling without getting wet?*

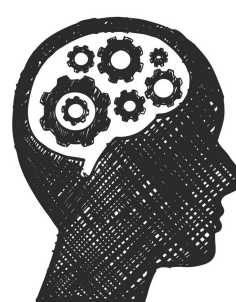
✗ **waterfall**, ✓ **bridge**, ✗ **valley**, ✗ **stream**, ✗ **bottom**

Claim: **Harry Potter** can teach classes on how to fly on a broomstick.



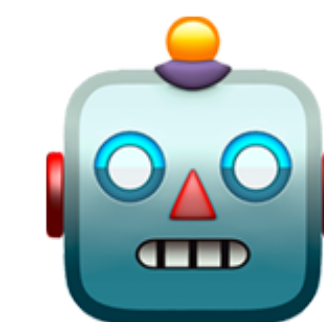
Harry Potter is a wizard ...
He plays Quidditch while riding on a broomstick.

+




Someone who's good at something can teach it.

True



Solving newer tasks requires...

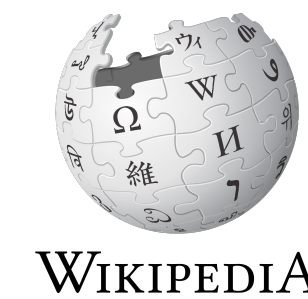


Sancho Panza is a character in Don Quixote.

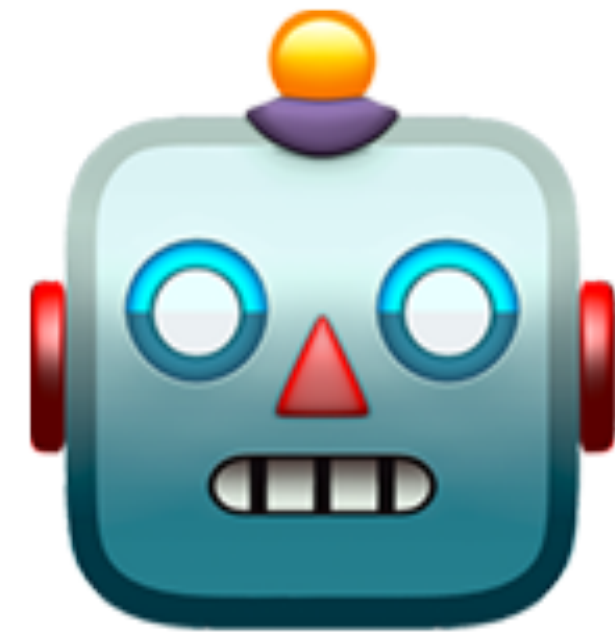
Damon Albarn's debut album was released in 2011.

He (Don Quixote) recruits a simple farmer, Sancho Panza, as his squire.

His (Damon Albarn's) debut solo studio album Everyday Robots was released in 2014.



entity knowledge



How much models know about entities?

Closed-Book (BART)



I remember 🧠!
The answer is _____

Retrieval-Based (DPR)



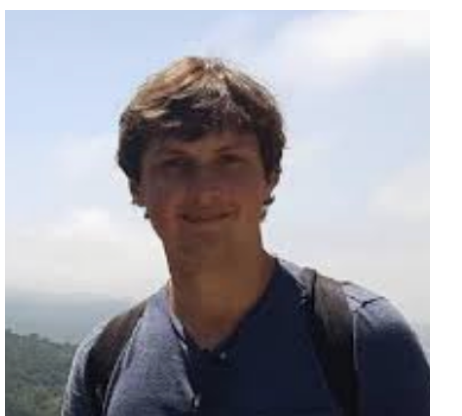
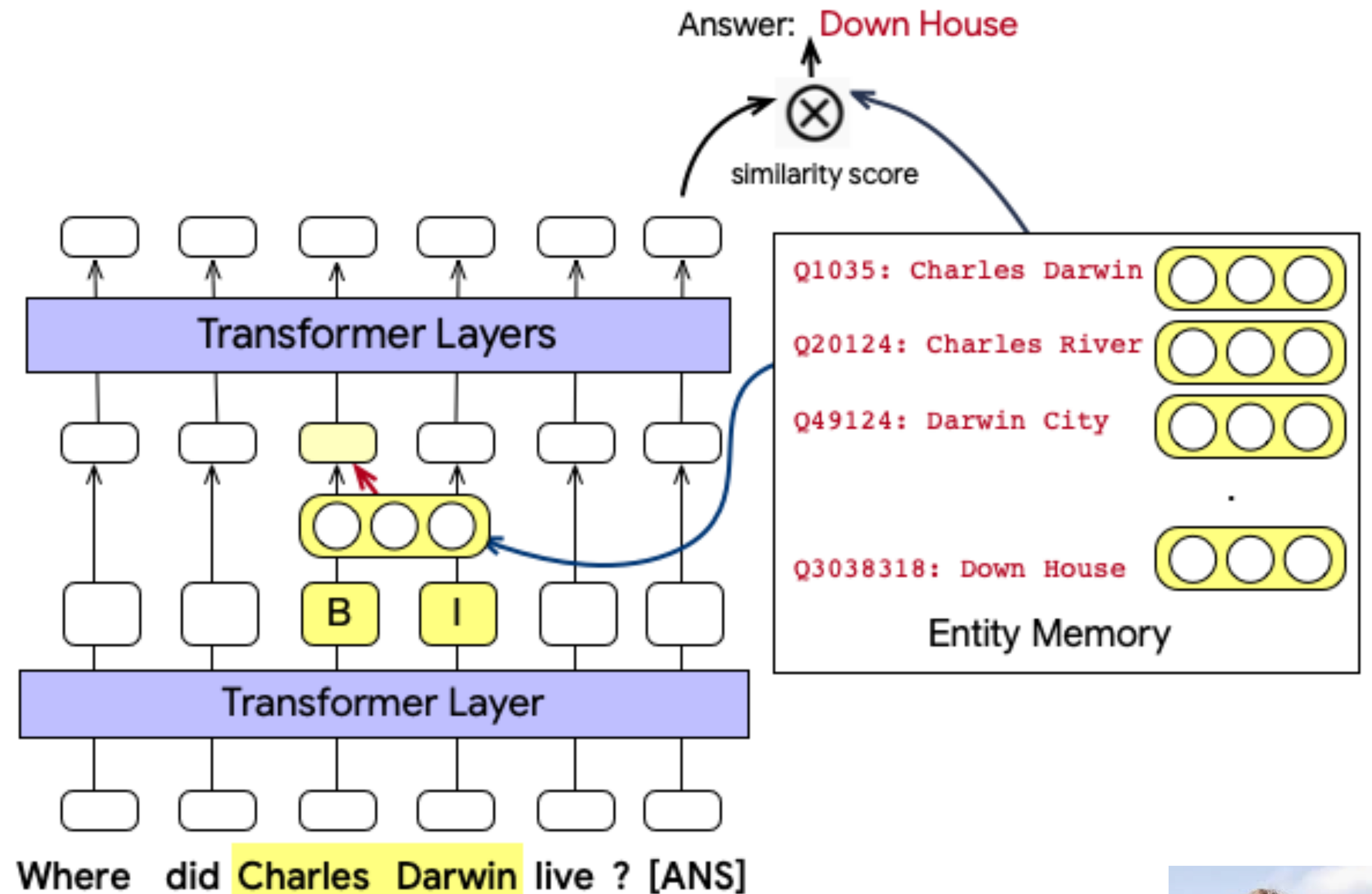
You found me 🕵️!
The answer is _____

- Both closed book approaches and retrieval based approaches perform competitively for filling in unambiguous facts about entities

Injecting Entity Knowledge into LMs 🧠

Predict answer entities from entity vocabulary

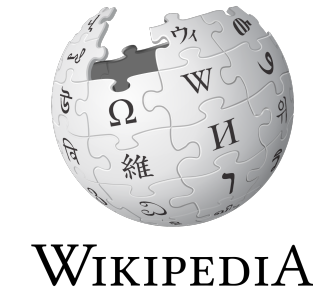
Construct a vector representation for entity, integrate it into query encoding



Questions can require implicit reasoning

Sancho Panza is a character in Don Quixote.

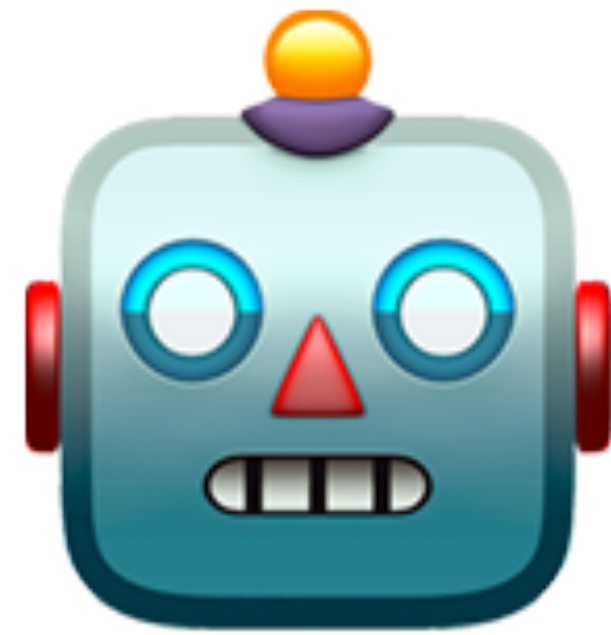
He (Don Quixote) recruits a simple farmer, Sancho Panza, as his squire.



entity knowledge

Damon Albarn's debut album was released in 2011.

His (Damon Albarn's) debut solo studio album Everyday Robots was released in 2014.



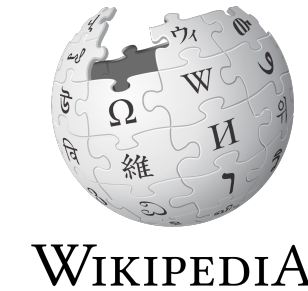
Can Aristotle use a laptop?
[Geva et al, TACL 2021]

????

Questions can require implicit reasoning

Sancho Panza is a character in Don Quixote.

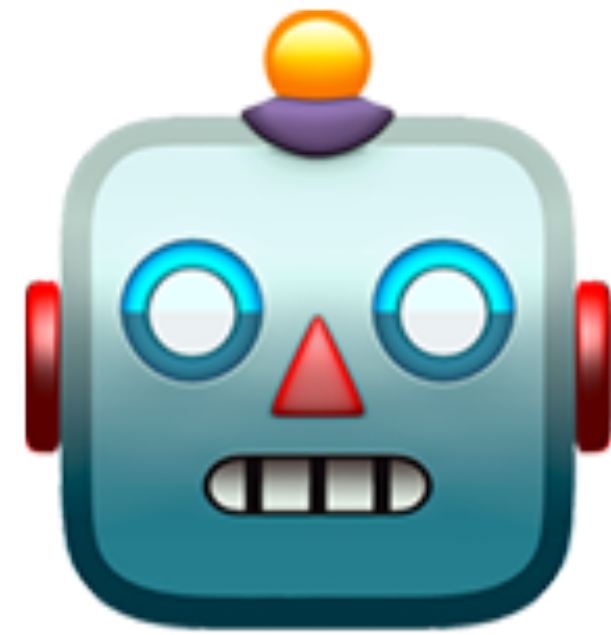
He (Don Quixote) recruits a simple farmer, Sancho Panza, as his squire.



entity knowledge

Damon Albarn's debut album was released in 2011.

His (Damon Albarn's) debut solo studio album Everyday Robots was released in 2014.



Can Aristotle use a laptop?
[Geva et al, TACL 2021]

????

One can drive from La Jolla to NYC in less than two hours.

????

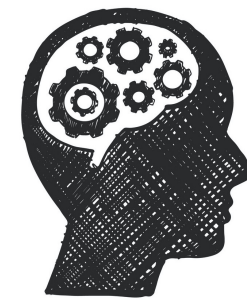
CREAK: A Dataset for Commonsense Reasoning over Entity Knowledge

- Total **13K** crowdsourced statements (half false, half true) covering 2K entities

Claim: Harry Potter can teach classes on how to fly on a broomstick.

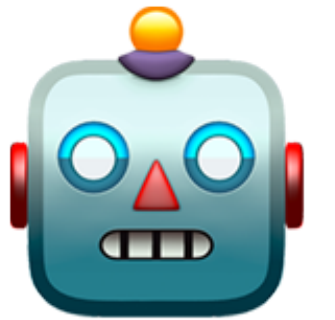


Harry Potter is a wizard ...
He plays Quidditch while riding on
a broomstick.

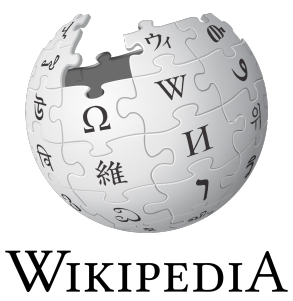


Someone who's good at
something can teach it.

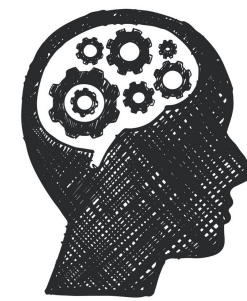
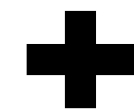
True



Claim: One can drive La Jolla to New York City in less than two hours.

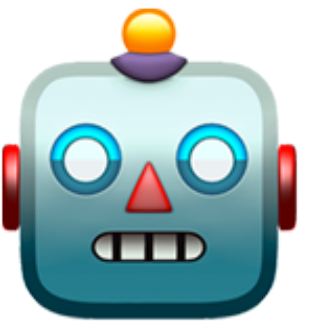


La Jolla is in California.
NYC is in New York.

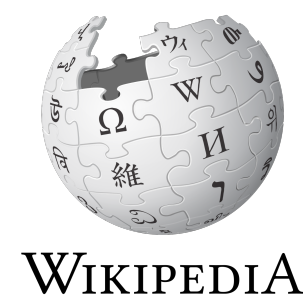


It takes 5h with airplane to fly from
California to New York.

False

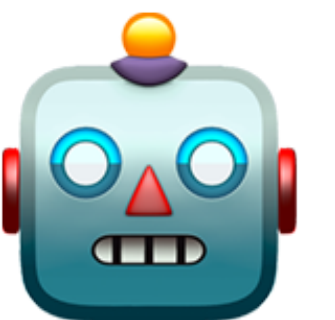


Claim: François Mitterrand became a Texas Senator in 2001.



François Mitterrand (26 Oct 1916
– 8 Jan 1996) was a French
statesman.

False



Manual Analysis: What's required to solve this task?

Retrieval (18%)

- We can find a very relevant sentence in the retrieval corpus.
- e.g., *The Harry Potter series originally began with the books.*

Common sense (28%)

- Complex reasoning but easily verifiable
- e.g., *Only trumpet players can perform a solo.*

A mix of retrieval and common sense (54%)

- e.g., *One can drive from La Jolla to New York City in less than two hours.*

Manual Analysis: What's required to solve this task?

Retrieval (18%)

- We can find a very relevant sentence in the retrieval corpus.
- e.g., *The Harry Potter series originally began with the books.*

Common sense (28%)

- Complex reasoning but easily verifiable
- e.g., *Only trumpet players can perform a solo.*

A mix of retrieval and common sense (54%)

- e.g., *One can drive from La Jolla to New York City in less than two hours.*

- Right/wrong affiliation/classification
- Similar sounds (e.g., dessert and desert)
- Matching/mismatching sense organs and their inputs (e.g., measuring a foot by ears)
- Malfunctioning system (e.g., human without a heart)
- Right/wrong characteristic (e.g., helium is the lightest element)
- Inability (e.g., a human can safely eat plutonium, Justin Bieber performed at Abraham Lincoln's inauguration)
 - Physically impossible
 - Geographically impossible
 - Temporally impossible
- Imaginary event
- Negating a true statement
- Matching/mismatching position/location (e.g., pancreas is right above the brain)
- Matching/mismatching tradition (e.g., camels are common pets in US)
- Matching/mismatching temporal information (e.g., Easter in October)
- Matching/mismatching functionality
- Matching/mismatching purpose
- Outdated information (i.e., not true anymore)

Manual Analysis: What's required to solve this task?

Retrieval (18%)

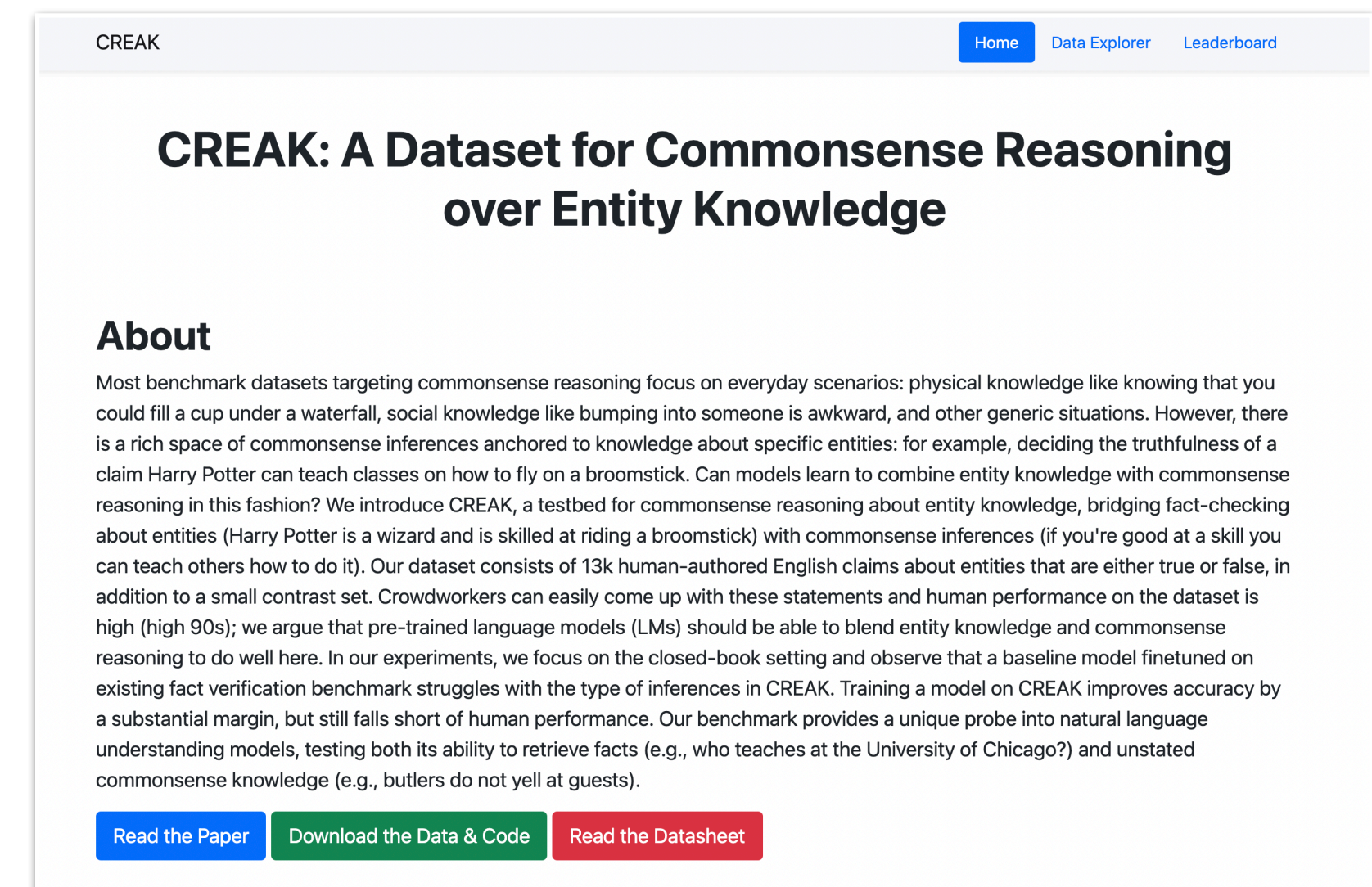
- We can find a very relevant sentence in the retrieval corpus.
- e.g., *The Harry Potter series originally began with the books.*

Common sense (28%)

- Complex reasoning but easily verifiable
- e.g., *Only trumpet players can perform a solo.*

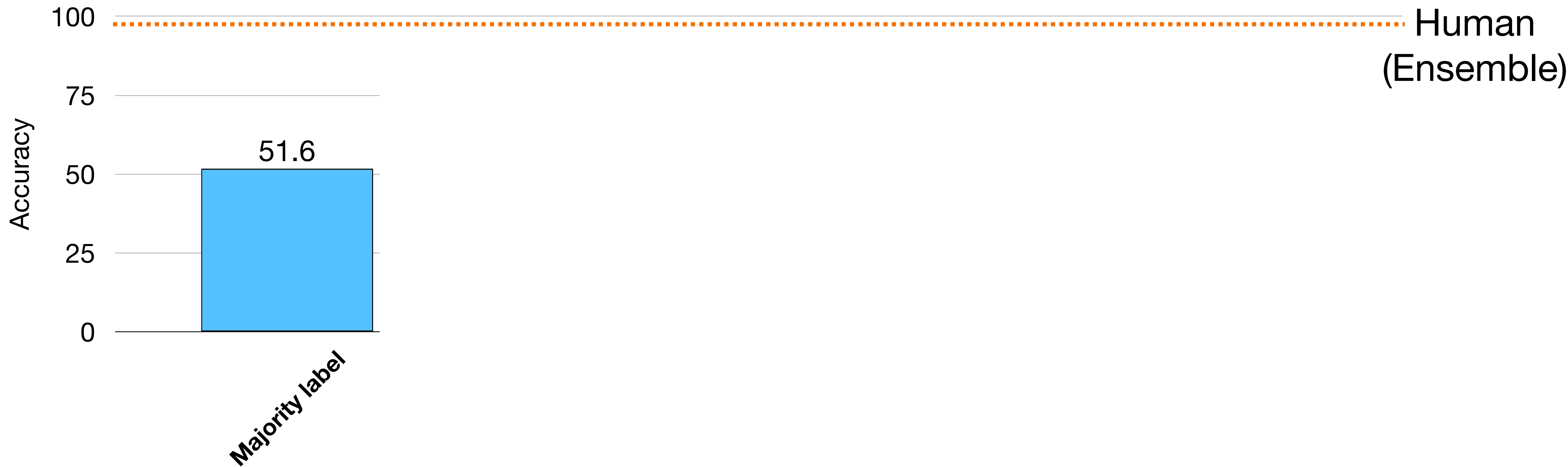
A mix of retrieval and common sense (54%)

- e.g., *One can drive from La Jolla to New York City in less than two hours.*

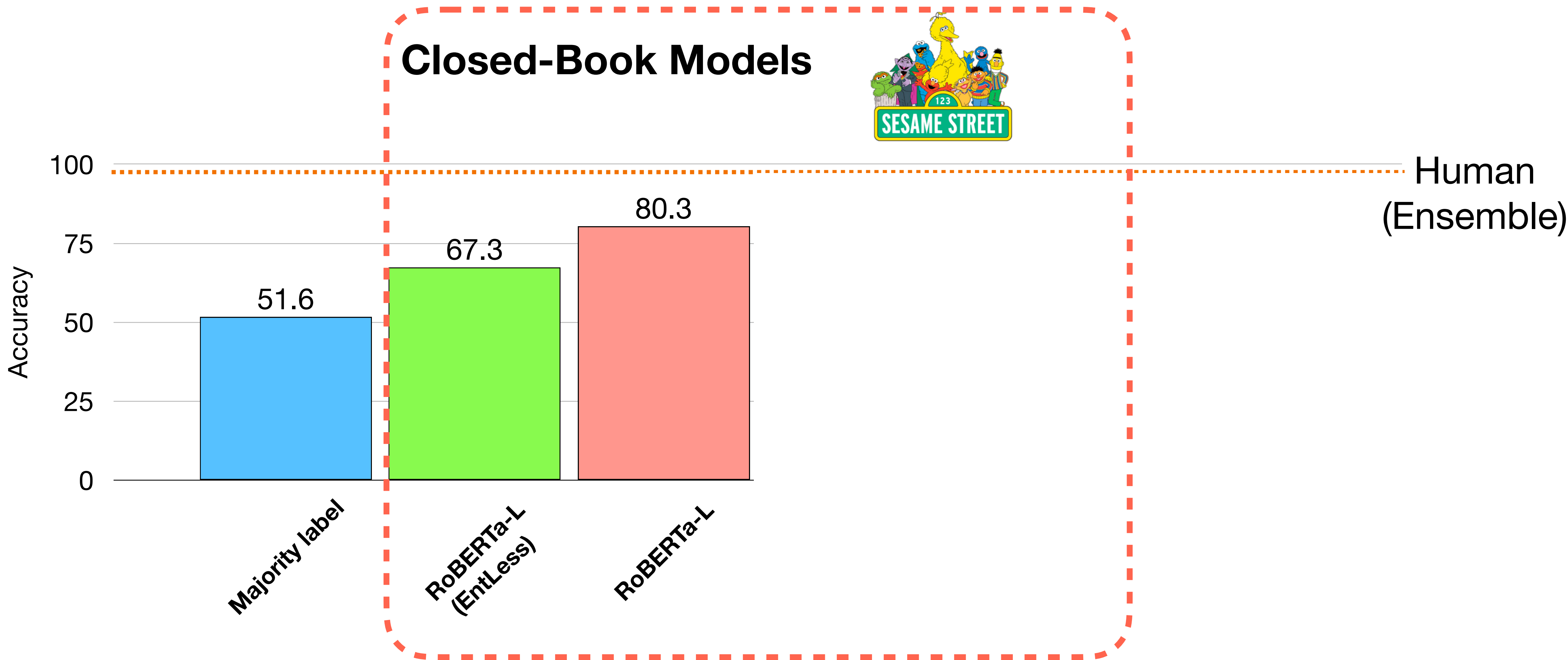


<https://www.cs.utexas.edu/~yasumasa/creak/>

Experimental Results



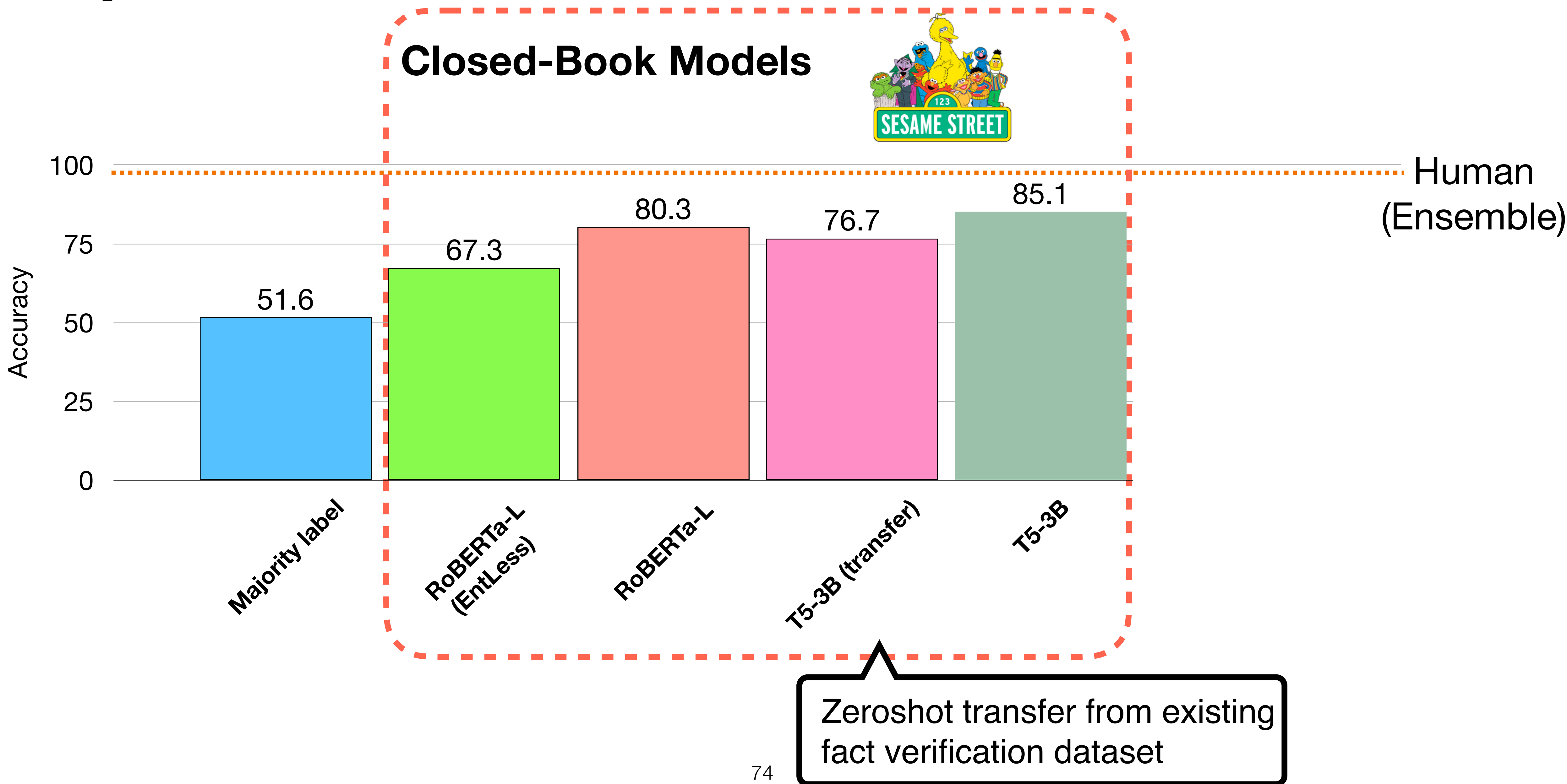
Experimental Results



Experimental Results



Experimental Results

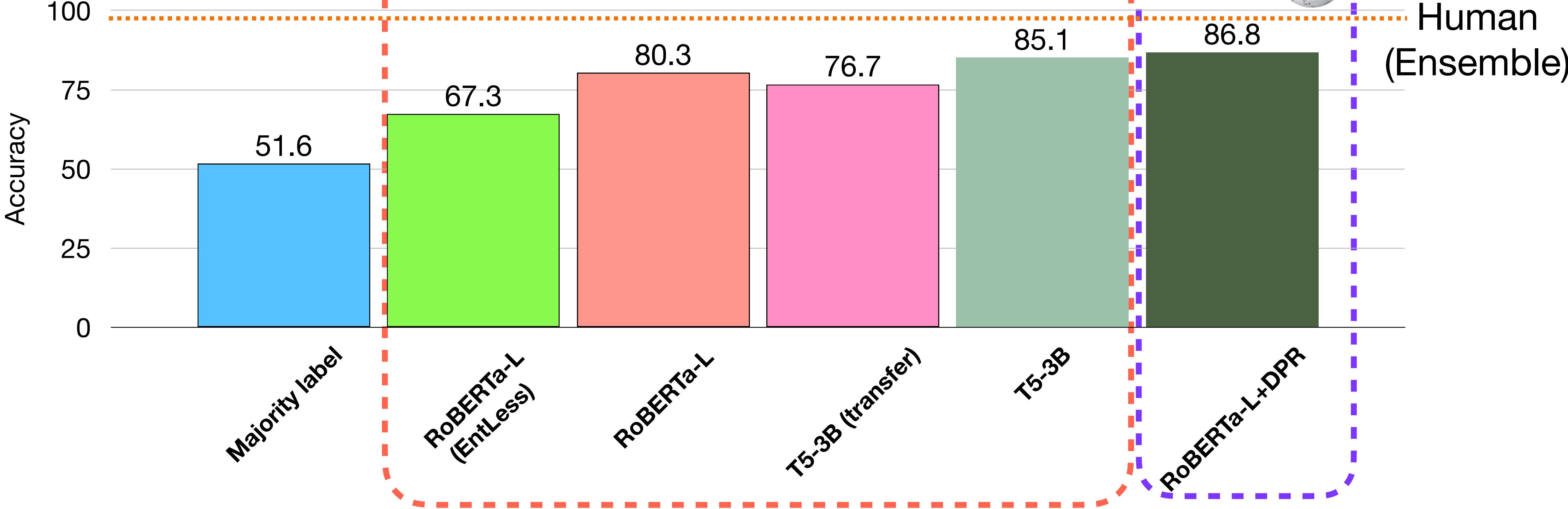


Experimental Results

Closed-Book Models



Retrieval-Based (DPR)



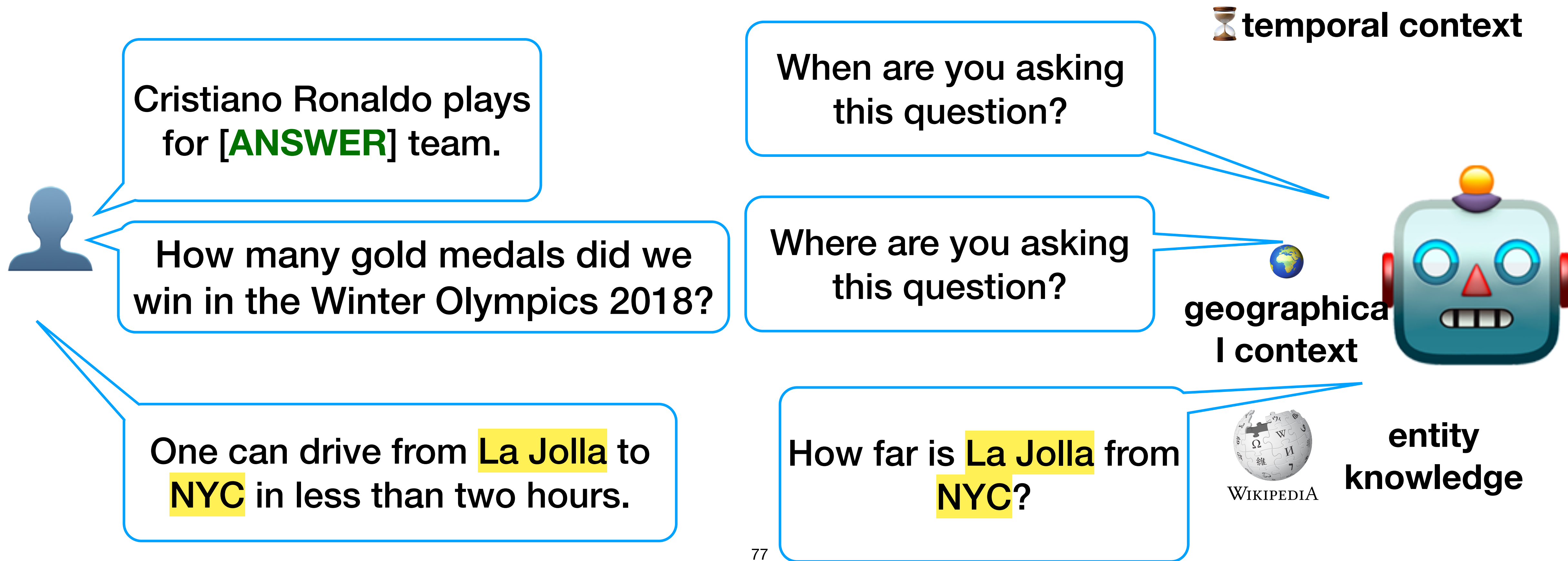
Summary

CREAK aims to evaluate models' ability to do reasoning based on facts about real world entities.

Bigger models perform better, retrieval helps somewhat, still lags (10+ points) behind human performance even ***without*** adversarial filtering

Today

- Understanding the extra-linguistic contexts
- Reasoning based on facts about entities



Open Questions

- **Understanding the extra-linguistic contexts**
 - How can we keep our benchmarks up-to-date?
 - How can we adapt models to rarer, newer contexts?
 - What other extra-linguistic contexts should consider?
- **Reasoning based on facts about entities**
 - How can we inject new information about entities to model?
 - Understanding model's failures — fact memorization / retrieval vs. reasoning

Thank You!

The screenshot shows the homepage of the SituatedQA project. At the top, there are navigation links for 'Home', 'Data Explorer', and 'Leaderboard'. The main heading is 'SituatedQA: Incorporating Extra-Linguistic Contexts into QA'. Below this is an 'About' section. To the right, there is a diagram illustrating a temporal context. A timeline shows dates from 12/2020 to 05/2021. A red box labeled 'Previous Answer: Moderna, Pfizer' is associated with the period from 12/2020 to 02/2021. A green box labeled 'Current Answer: Moderna, Pfizer, J&J' is associated with the period from 03/2021 to 05/2021. Below the timeline, two boxes show context and answer pairs: one for Dec 18, 2014 (Answer: Moderna, Pfizer) and one for Apr 10, 2021 (Answer: Moderna, Pfizer, J&J).

SituatedQA Home Data Explorer Leaderboard

SituatedQA: Incorporating Extra-Linguistic Contexts into QA

About

Answers to the same question may change depending on the extra-linguistic contexts (when and where the question was asked). To study this challenge, we introduce **SituatedQA**, an open-retrieval QA dataset where systems must produce the correct answer to a question given the temporal or geographical context. To construct **SituatedQA** we first identify such questions in existing QA datasets. We find that a significant proportion of information seeking questions have context-dependent answers (e.g., roughly 16.5% of NQ-Open). For such context-dependent questions, we then crowdsource alternative contexts and their corresponding answers. Our study shows that existing models

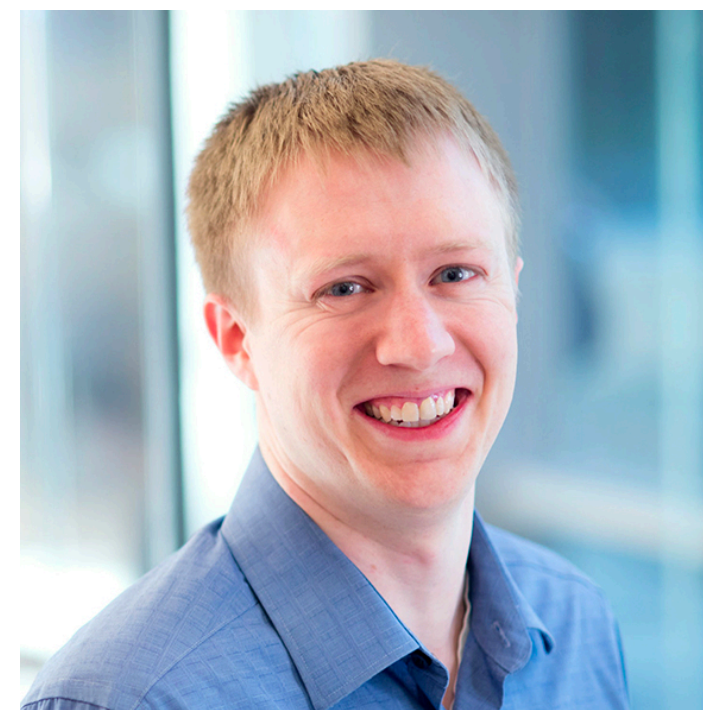
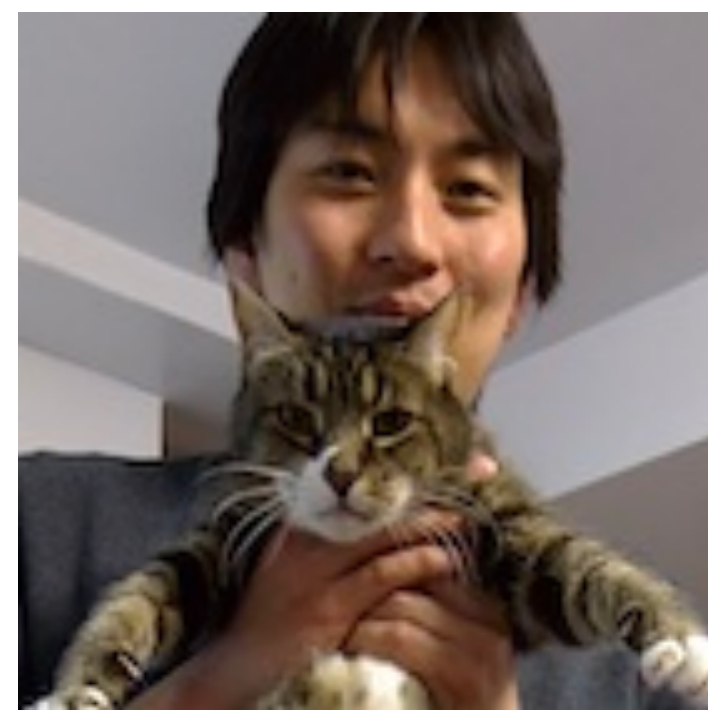
Context Type: Temporal
Question: Which COVID-19 vaccines have been authorized for adults in the US?

Previous Answer: Moderna, Pfizer
Current Answer: Moderna, Pfizer, J&J

Context: Dec 18, 2014
Answer: Moderna, Pfizer

Context: Apr 10, 2021
Answer: Moderna, Pfizer, J&J

Website: <https://situatedqa.github.io/>



The screenshot shows the homepage of the CREAK project. At the top, there are navigation links for 'Home', 'Data Explorer', and 'Leaderboard'. The main heading is 'CREAK: A Dataset for Commonsense Reasoning over Entity Knowledge'. Below this is an 'About' section. At the bottom, there are three buttons: 'Read the Paper', 'Download the Data & Code', and 'Read the Datasheet'.

CREAK Home Data Explorer Leaderboard

CREAK: A Dataset for Commonsense Reasoning over Entity Knowledge

About

Most benchmark datasets targeting commonsense reasoning focus on everyday scenarios: physical knowledge like knowing that you could fill a cup under a waterfall, social knowledge like bumping into someone is awkward, and other generic situations. However, there is a rich space of commonsense inferences anchored to knowledge about specific entities: for example, deciding the truthfulness of a claim Harry Potter can teach classes on how to fly on a broomstick. Can models learn to combine entity knowledge with commonsense reasoning in this fashion? We introduce CREAK, a testbed for commonsense reasoning about entity knowledge, bridging fact-checking about entities (Harry Potter is a wizard and is skilled at riding a broomstick) with commonsense inferences (if you're good at a skill you can teach others how to do it). Our dataset consists of 13k human-authored English claims about entities that are either true or false, in addition to a small contrast set. Crowdworkers can easily come up with these statements and human performance on the dataset is high (high 90s); we argue that pre-trained language models (LMs) should be able to blend entity knowledge and commonsense reasoning to do well here. In our experiments, we focus on the closed-book setting and observe that a baseline model finetuned on existing fact verification benchmark struggles with the type of inferences in CREAK. Training a model on CREAK improves accuracy by a substantial margin, but still falls short of human performance. Our benchmark provides a unique probe into natural language understanding models, testing both its ability to retrieve facts (e.g., who teaches at the University of Chicago?) and unstated commonsense knowledge (e.g., butlers do not yell at guests).

[Read the Paper](#) [Download the Data & Code](#) [Read the Datasheet](#)

Website: <https://www.cs.utexas.edu/~yasumasa/creak/>