

SituatedQA: Incorporating Extra-Linguistic Contexts into QA

Michael J.Q. Zhang and Eunsol Choi



Motivation

Where were the last Winter Olympic Games held?



List of Winter Olympic Games

No. ⚡	Year ⚡	Host ⚡
XXII	2014	 Sochi
XXIII	2018	 Pyeongchang
XXIV	2022	 Beijing



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



2018 Winter Olympics Medal Table

Rank ⚡	NOC ▲	Gold ⚡
?	 Australia (AUS)	0
?	 Austria (AUT)	5
?	 Belarus (BLR)	2
?	 Belgium (BEL)	0

Motivation

Where were the last Winter Olympic Games held?

Sochi, Russia

List of Winter Olympic Games

No. ⚡	Year ⚡	Host ⚡
XXII	2014	 Sochi
XXIII	2018	 Pyeongchang
XXIV	2022	 Beijing



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



Motivation

Where were the last Winter Olympic Games held?

Sochi, Russia

List of Winter Olympic Games

No. ⚡	Year ⚡	Host ⚡
XXII	2014	 Sochi
XXIII	2018	 Pyeongchang
XXIV	2022	 Beijing



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

5

2018 Winter Olympics Medal Table

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

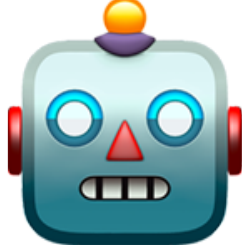



Motivation

Where were the last Winter Olympic Games held?

20% of answers depend on when the question was asked

Sochi, Russia



List of Winter Olympic Games

No. ⚡	Year ⚡	Host ⚡
XXII	2014	 Sochi
XXIII	2018	 Pyeongchang
XXIV	2022	 Beijing

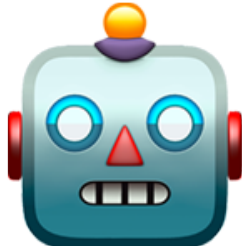

✗

✓


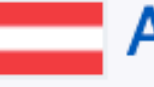


✗

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

5% of questions specify a location where it was asked



2018 Winter Olympics Medal Table

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

?

?

?


?

Motivation

Where were the last Winter Olympic Games held?

20% of answers depend on when the question was asked

List of Winter Olympic Games





No. ⚡	Year ⚡	Host ⚡
XXII	2014	 Sochi

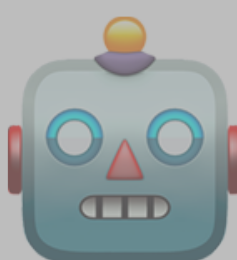
The answer to a question may **depend on its extra-linguistic contexts**
(e.g., **when** and **where** it was asked)

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

5% of questions specify a location where it was asked

Table

Rank ⚡	NOC	Gold ⚡
?	 Australia (AUS)	0
?	 Austria (AUT)	5
?	 Belarus (BLR)	2
?	 Belgium (BEL)	0



Motivation: Advances in QA Benchmarking

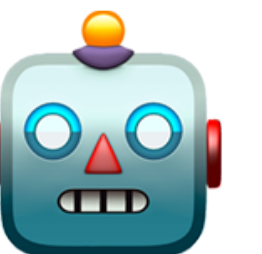


Where was the last Winter Olympic Games held?

Why is this an study issue *now*?

Recent advances in NLP have opened new doors for benchmarking QA models

PyeongChang



Motivation: Advances in QA Benchmarking



Where was the last Winter Olympic Games held?

Benchmarking in the Past

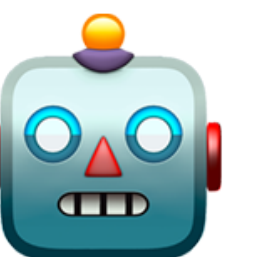
2018 Winter Olympics

The **2018 Winter Olympics**, officially known as the **XXIII Olympic Winter Games** (**French**: *Les XXIII^{es} Jeux olympiques d'hiver*,^[a] **Korean**: 제23회 동계 올림픽, **romanized**: *Jeisipsamhoe Donggye Ollimpik*) and commonly known as **PyeongChang 2018** (**Korean**: 평창 2018), was an international winter **multi-sport event** that was held between 9 and 25 February 2018 in **Pyeongchang** County, Gangwon Province, South Korea, with the opening rounds for certain events held on 8 February 2018, the day before the **opening ceremony**.

Past benchmarks assume:

- The answer is a **span** in the provided passage
- All the necessary context is given in the document

PyeongChang



Motivation: Advances in QA Benchmarking

Where was the last Winter Olympic Games held?

Benchmarking in the Past

2018 Winter Olympics

The **2018 Winter Olympics**, officially known as the **XXIII Olympic Winter Games** (**French**: *Les XXIII^{es} Jeux olympiques d'hiver*;^[a] **Korean**: 제23회 동계 올림픽, *romanized*: *Jeisipsamhoe Donggye Ollimpik*) and commonly known as **PyeongChang 2018** (**Korean**: 평창 2018), was an international winter **multi-sport event** that was held between 9 and 25 February 2018 in **Pyeongchang County, Gangwon Province, South Korea**, with the opening rounds for certain events held on 8 February 2018, the day before the **opening ceremony**.

Past benchmarks assume:

- The answer is a **span** in the provided passage
- All the necessary context is given in the document

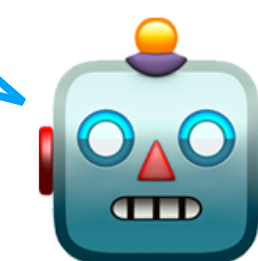
Benchmarking Today



Today's benchmarks assume:

- The answer is out there ***somewhere***...
- We have to assume the context

PyeongChang



The SituatedQA Task



Temporally Dependent Question

Where were the last Winter Olympic Games held?



Geographically Dependent Question

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



Objectives

1. Determine whether QA models understand how information changes across different contexts
2. Create QA systems that can produce the appropriate answer for a given context

The SituatedQA Task



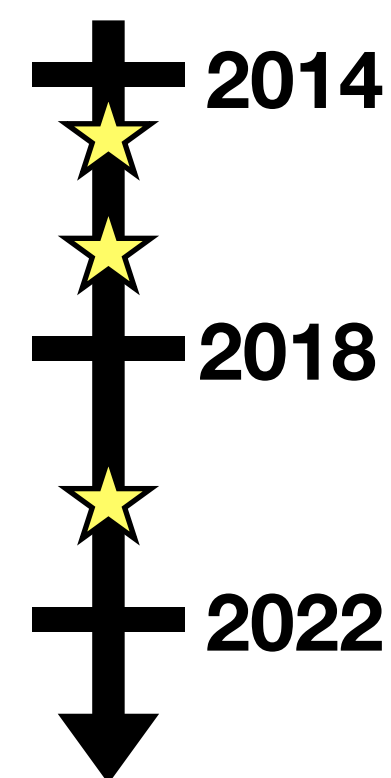
Temporally Dependent Question

Where were the last Winter Olympic Games held?



+

Temporal Context (Timestamp)



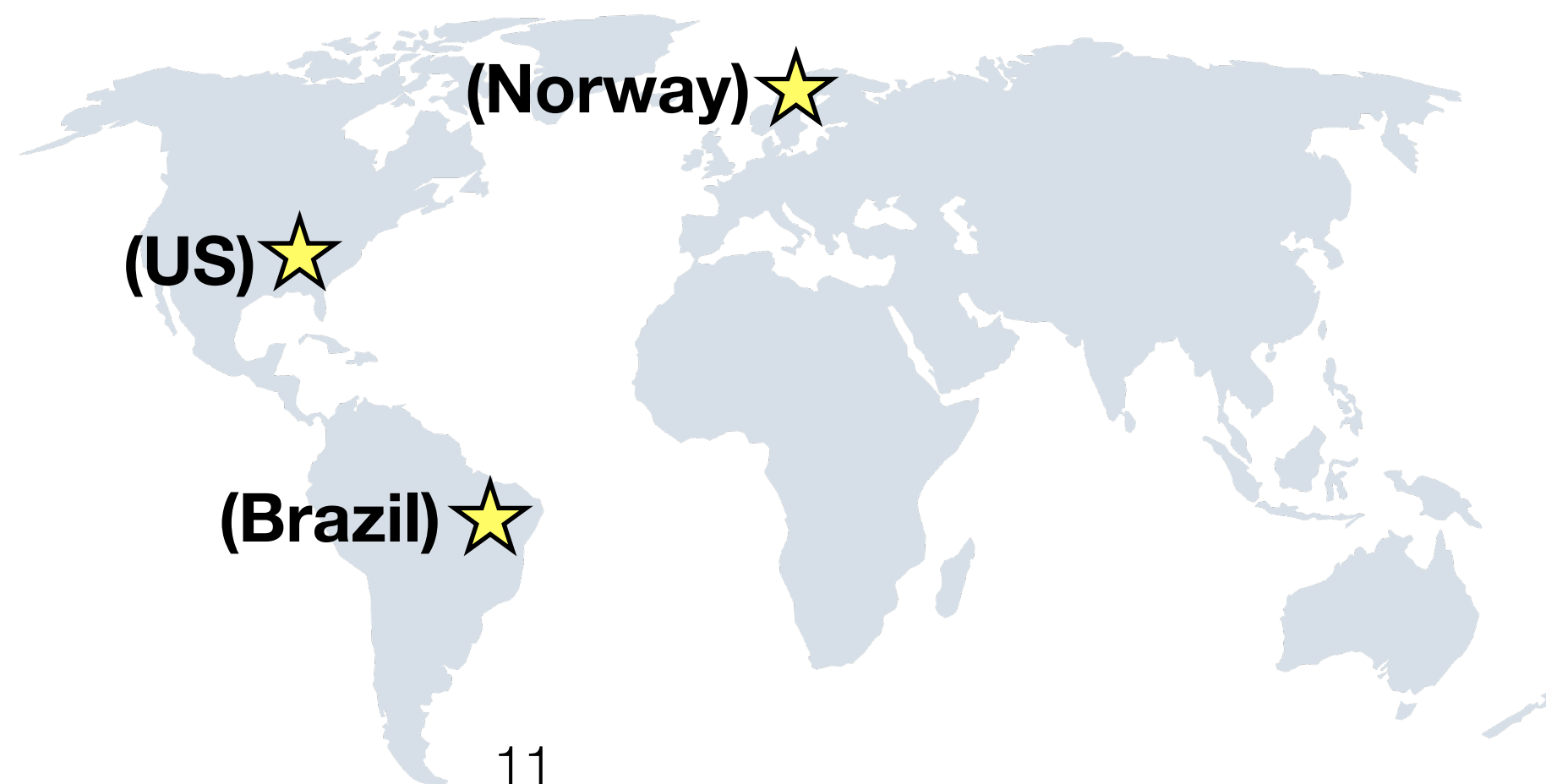
Geographically Dependent Question

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



+

Geographical Context (Location)



The SituatedQA Task



Temporally Dependent Question

Where were the last Winter Olympic Games held?

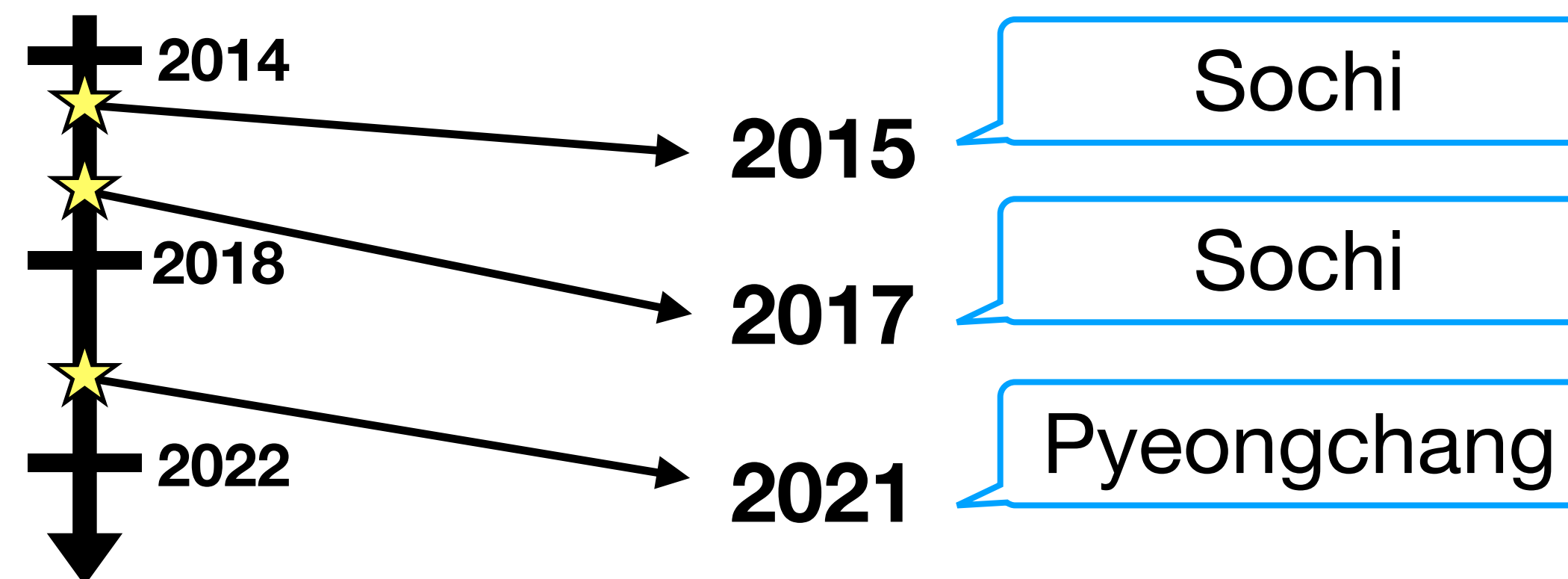


+

Temporal Context (Timestamp)



Answer



Geographically Dependent Question

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

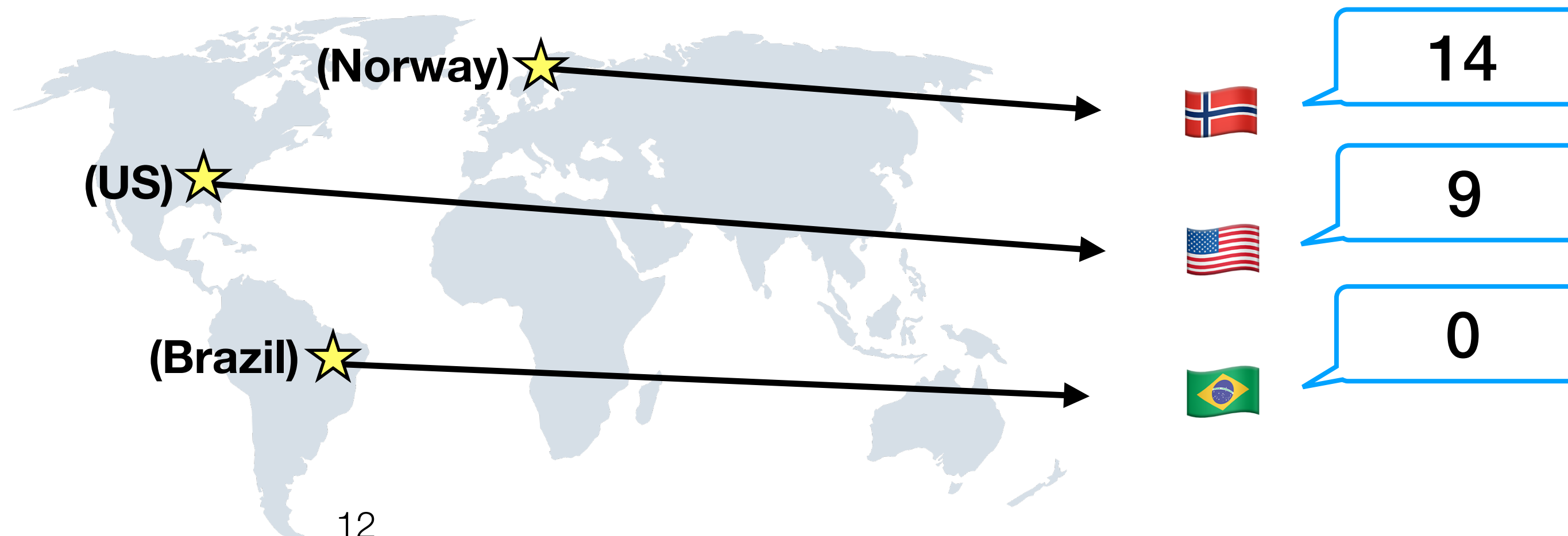


+

Geographical Context (Location)



Answer



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

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

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

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

We adopt a three stage pipeline for gathering annotations

0. Input Question

When was the last time states
were created?

Dataset: Geographical SituatedQA

We adopt a three stage pipeline for gathering annotations

0. Input Question

When was the last time states were created?

1. Identification

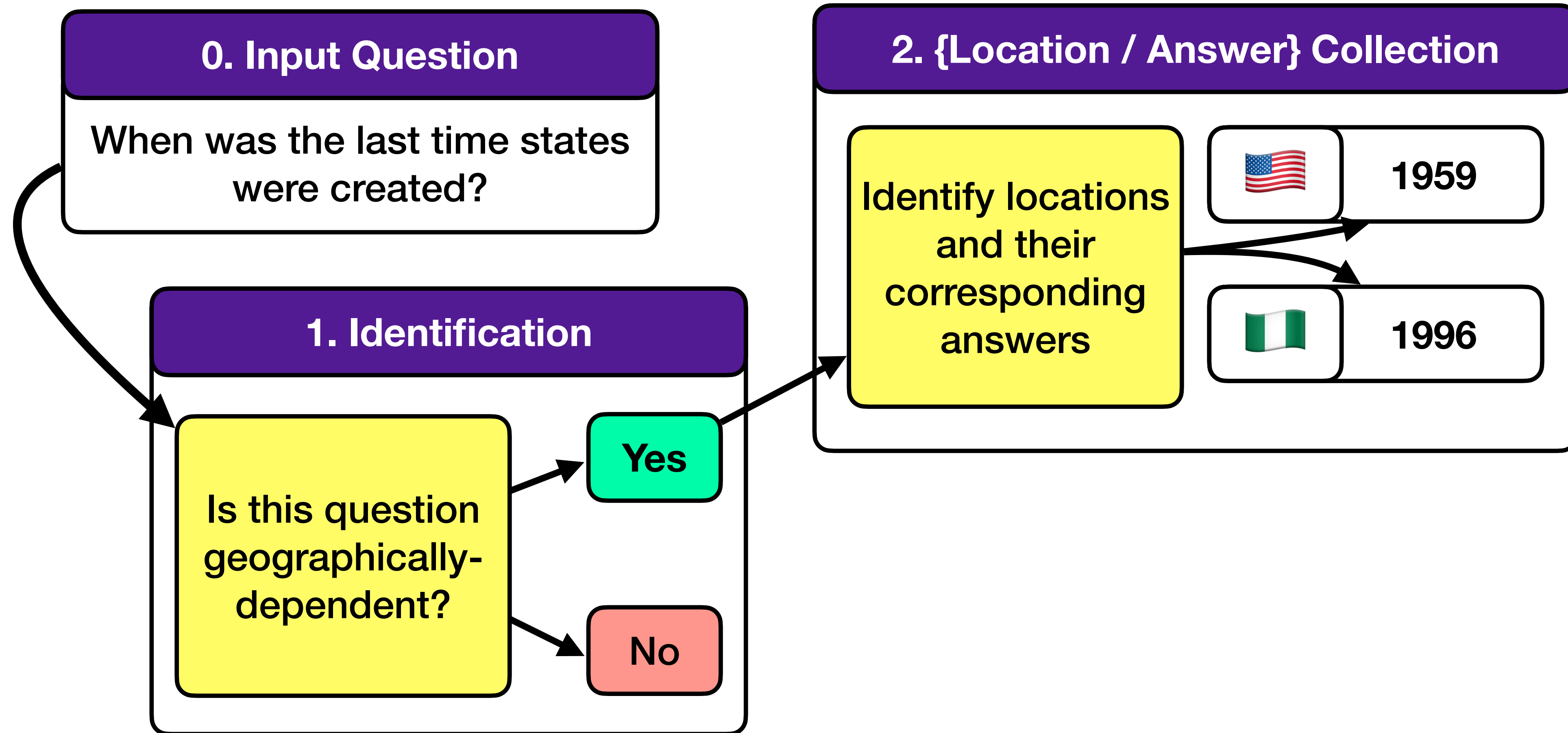
Is this question geographically-dependent?

Yes

No

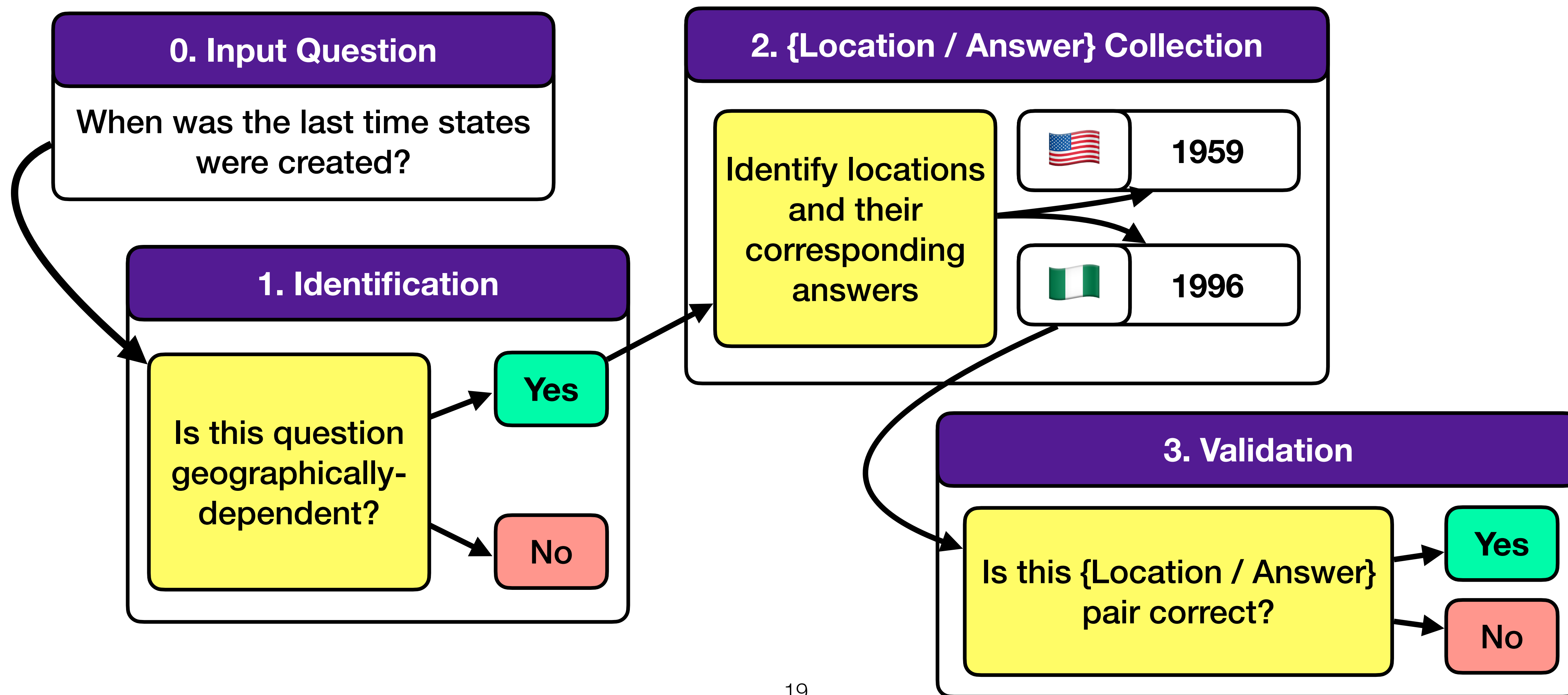
Dataset: Geographical SituatedQA

We adopt a three stage pipeline for gathering annotations



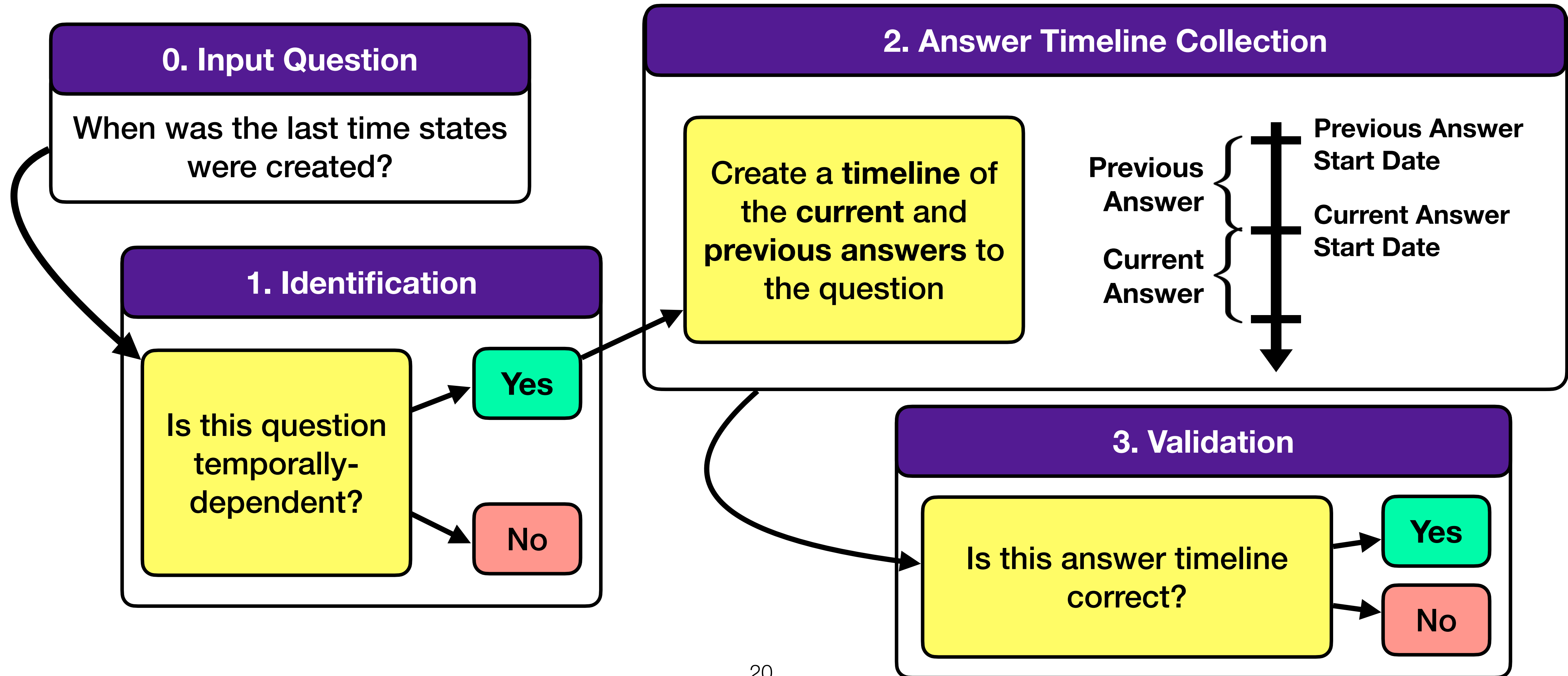
Dataset: Geographical SituatedQA

We adopt a three stage pipeline for gathering annotations



Dataset: Temporal SituatedQA

We adopt a three stage pipeline for gathering annotations

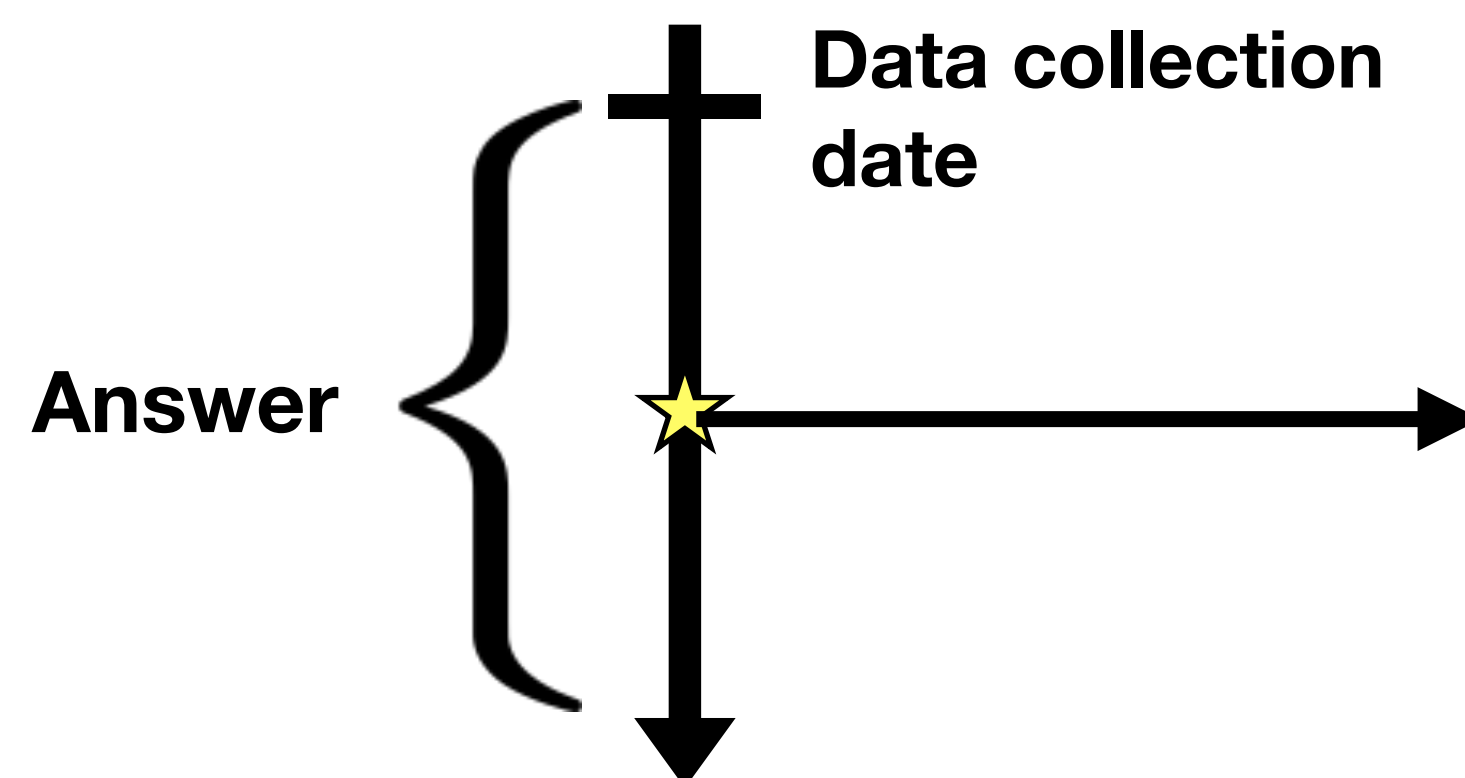


Mapping Answer Timelines to {Timestamps / Answers}



Mapping Answer Timelines to {Timestamps / Answers}

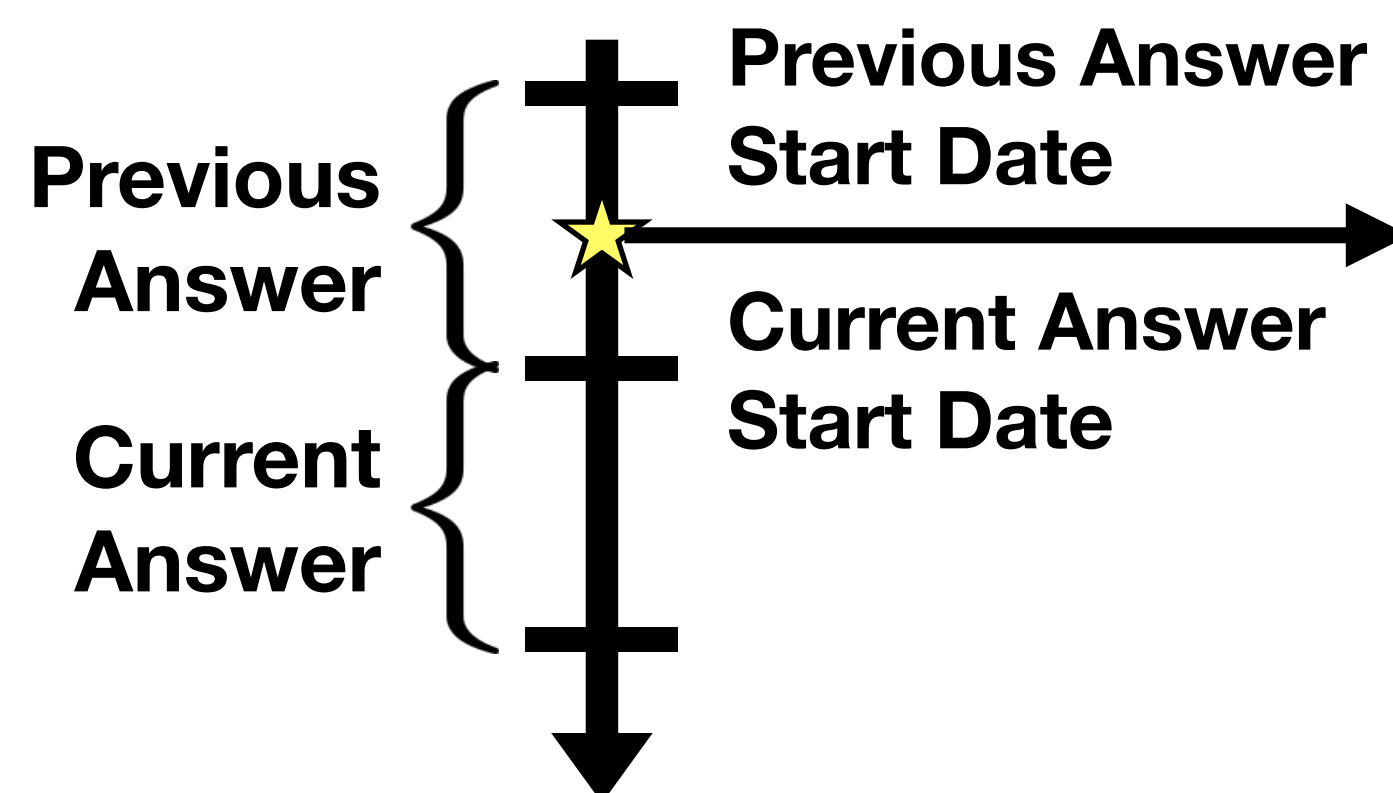
Static Questions



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

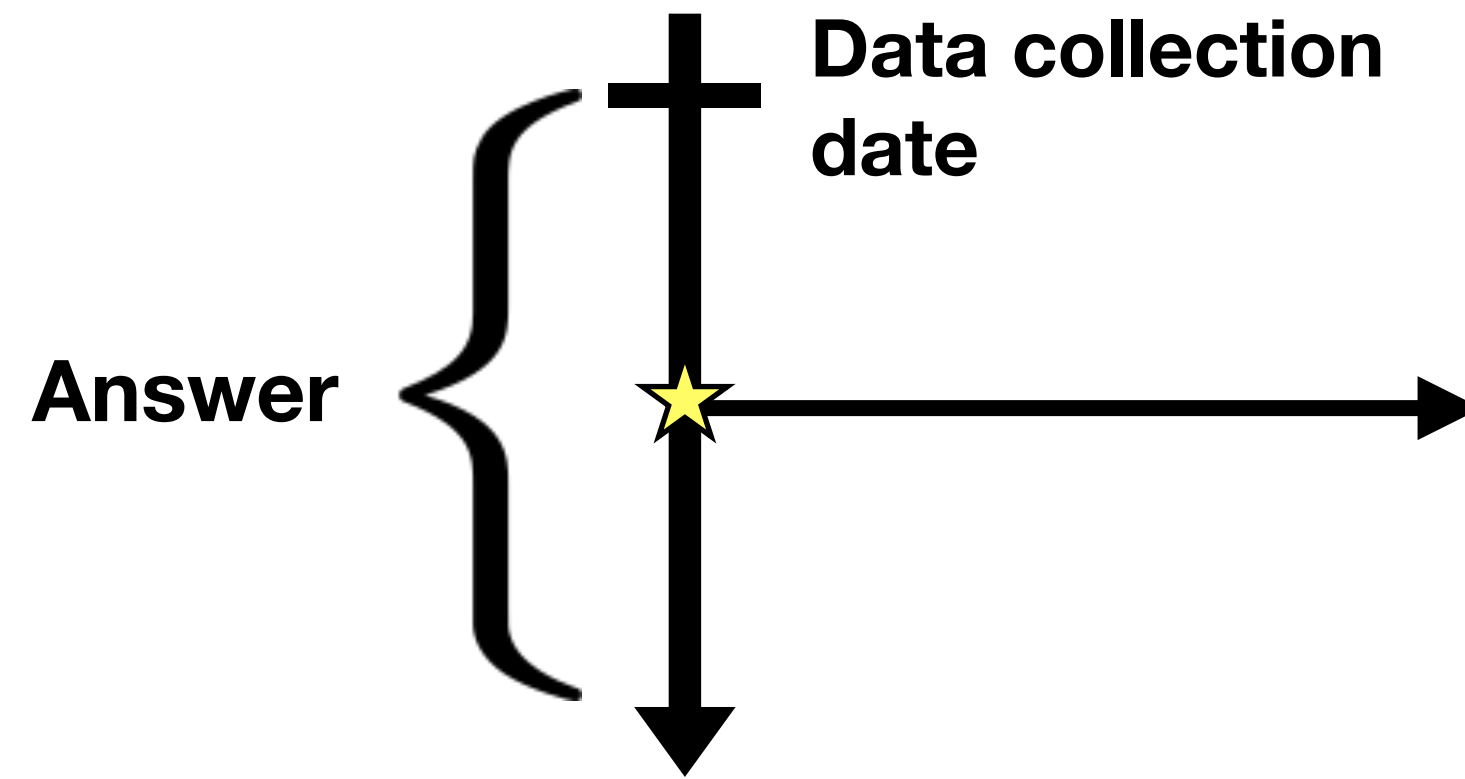


Sampled: selecting a timestamp between an answer's start and end date

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

Mapping Answer Timelines to {Timestamps / Answers}

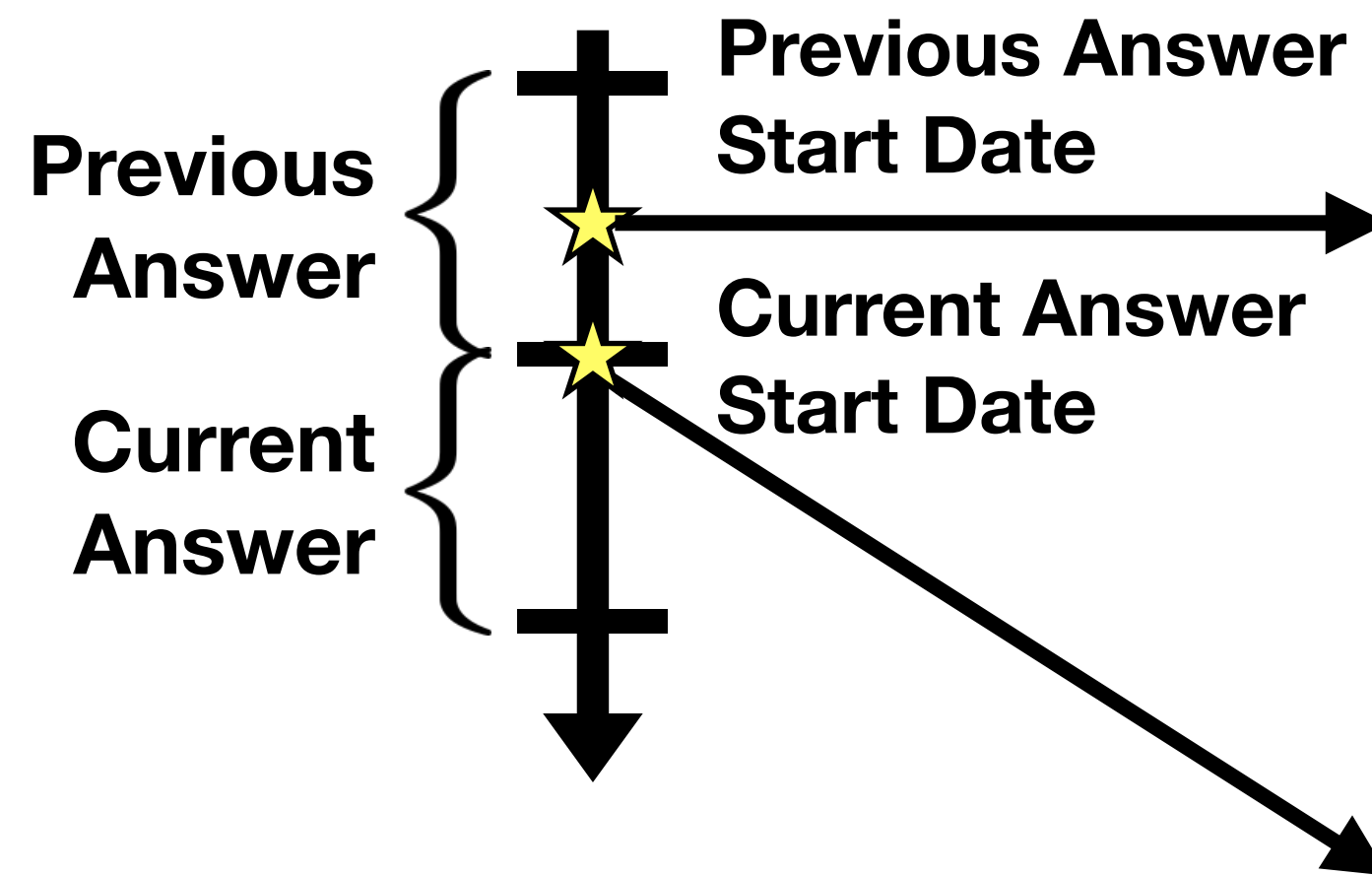
Static Questions



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



Temporally Dependent Question

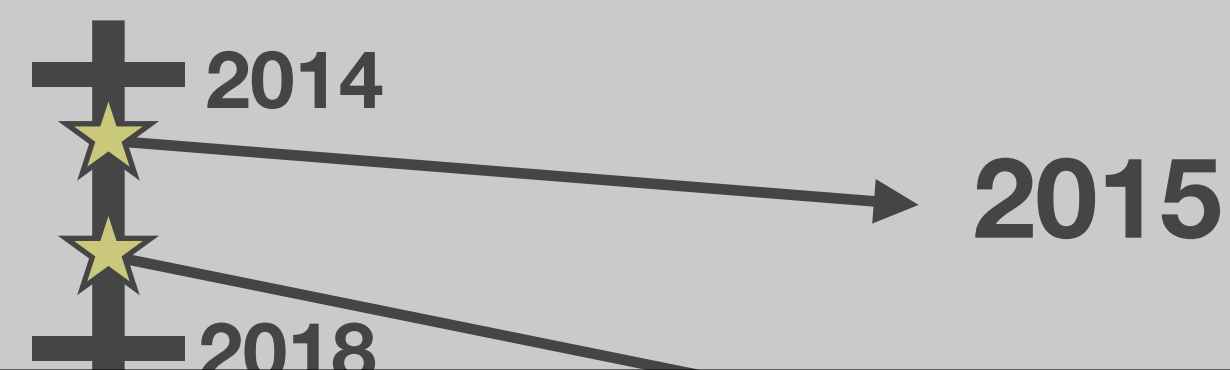
+

Temporal Context (Timestamp)



Answer

Where were the last Winter Olympic Games



Sochi

Sochi



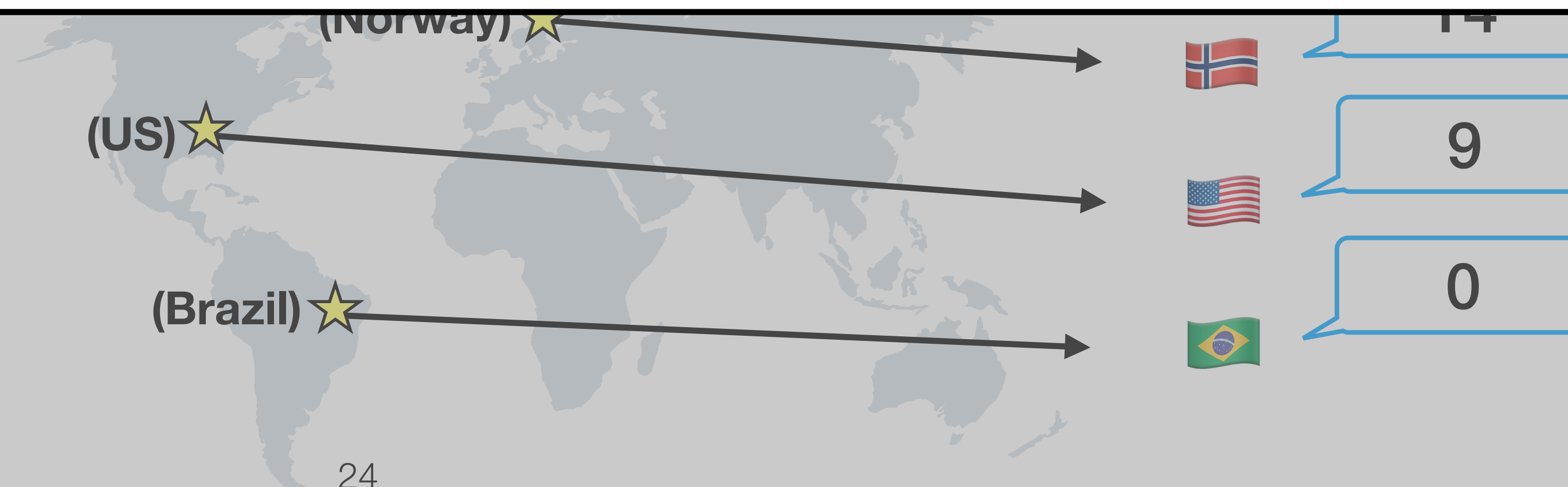
Sourcing from NQ-Open, we create:

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



G

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



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)

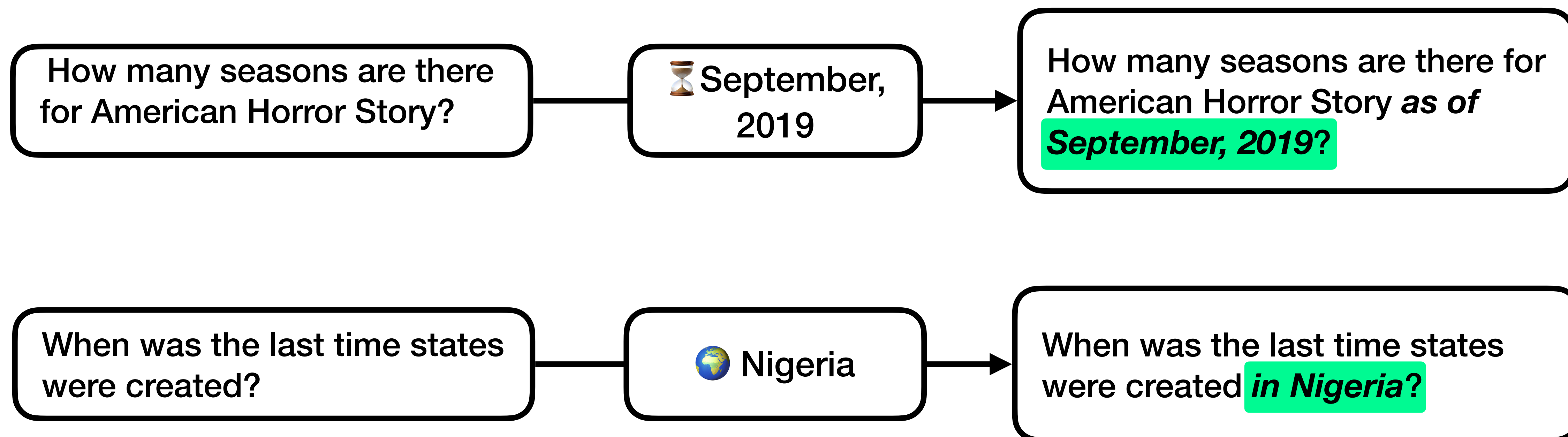


You found me 🕵️!
The answer is _____

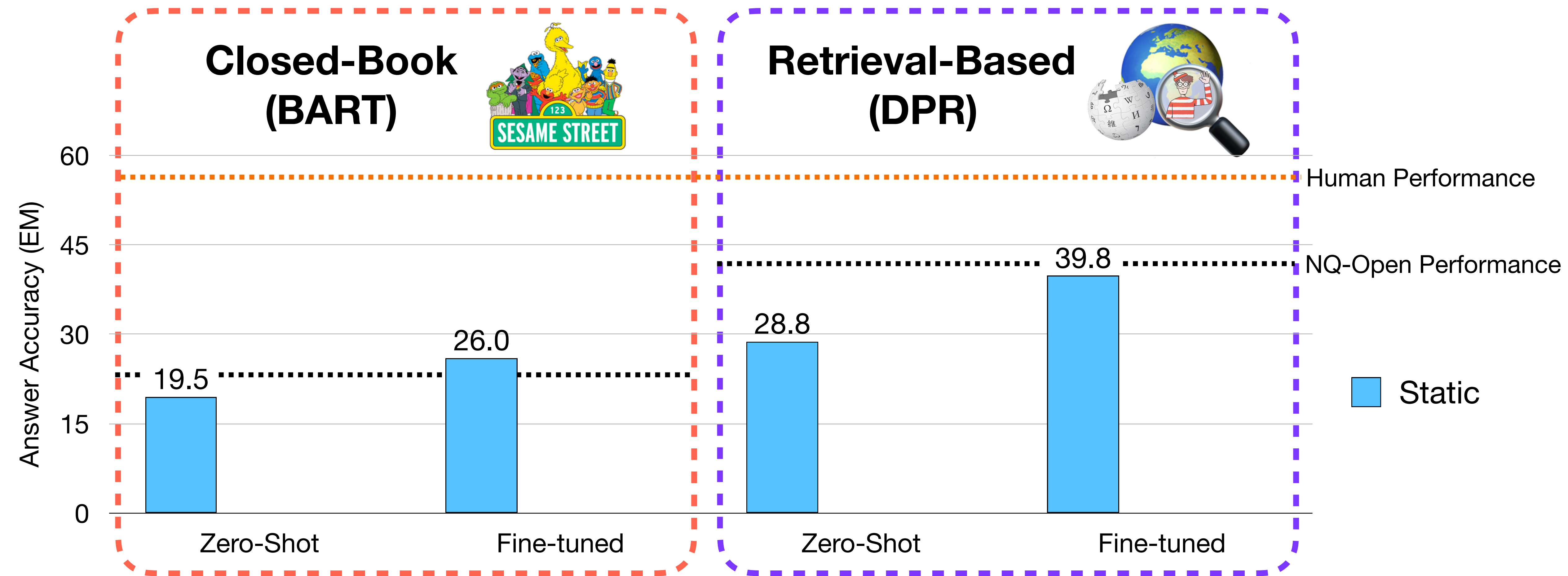
- **Zero-shot** models are trained on NQ-Open
- **Fine-tuned** models are first trained on NQ-Open, then also tuned on our collected training data

Experiments: Incorporating Context

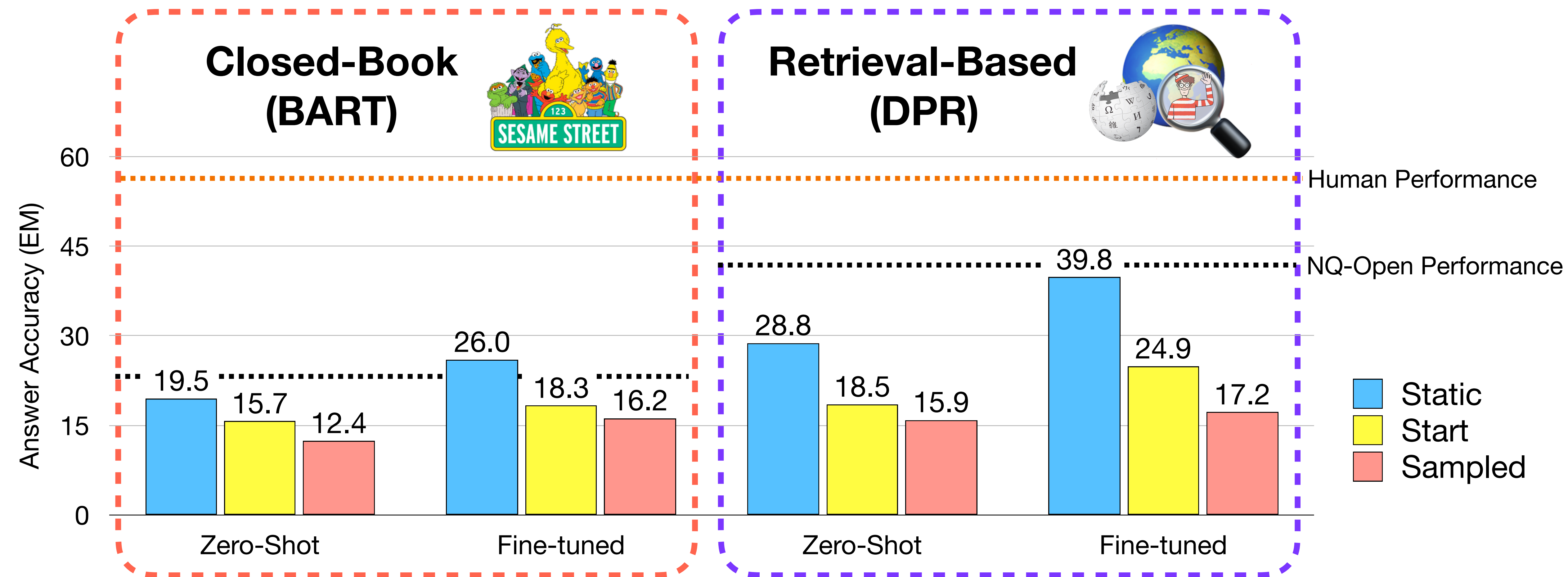
Query modification is used to specify extra-linguistic context to a model



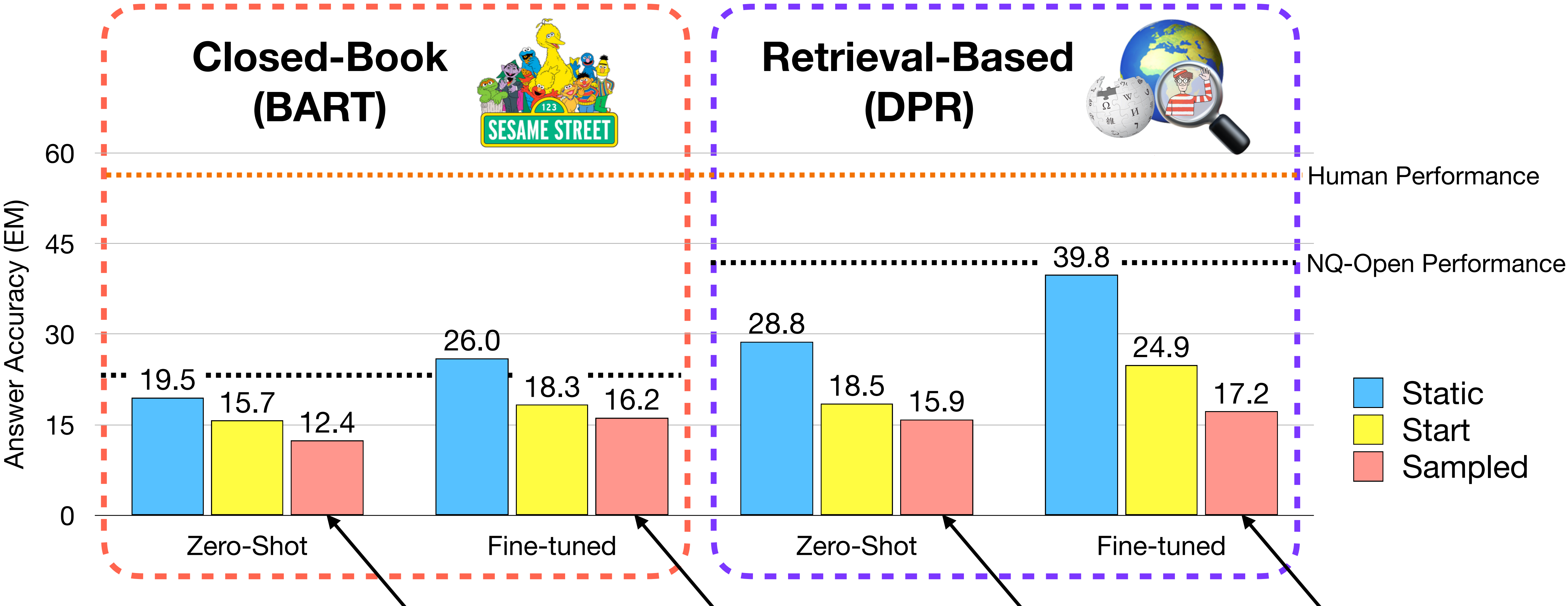
Results: Temporal SituatedQA



Results: Temporal SituatedQA



Results: Temporal SituatedQA



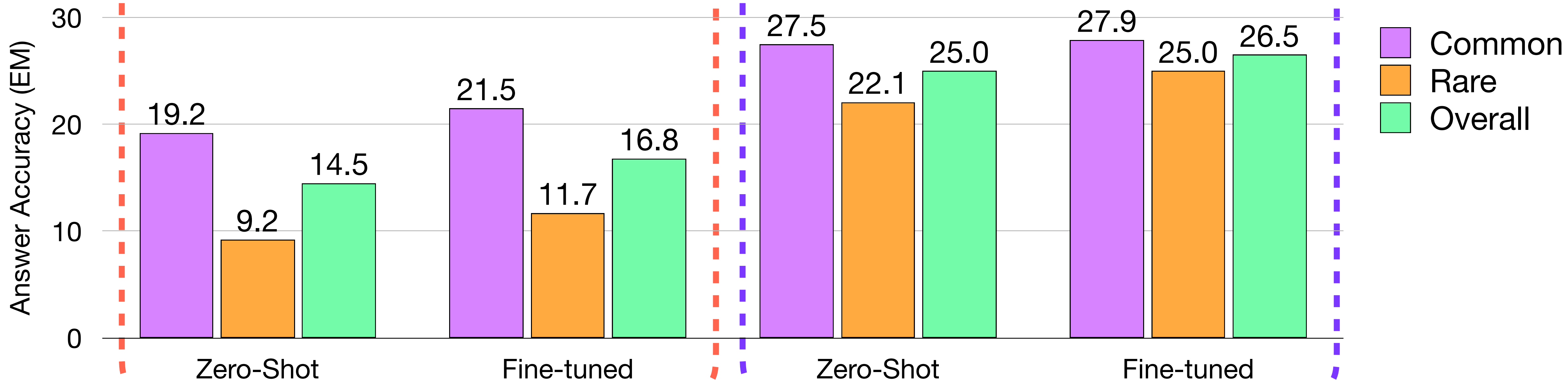
Models struggle the most with contexts that are *between* answer transitions

Results: Geographical SituatedQA

Closed-Book (BART)



Retrieval-Based (DPR)



Results: Geographical SituatedQA

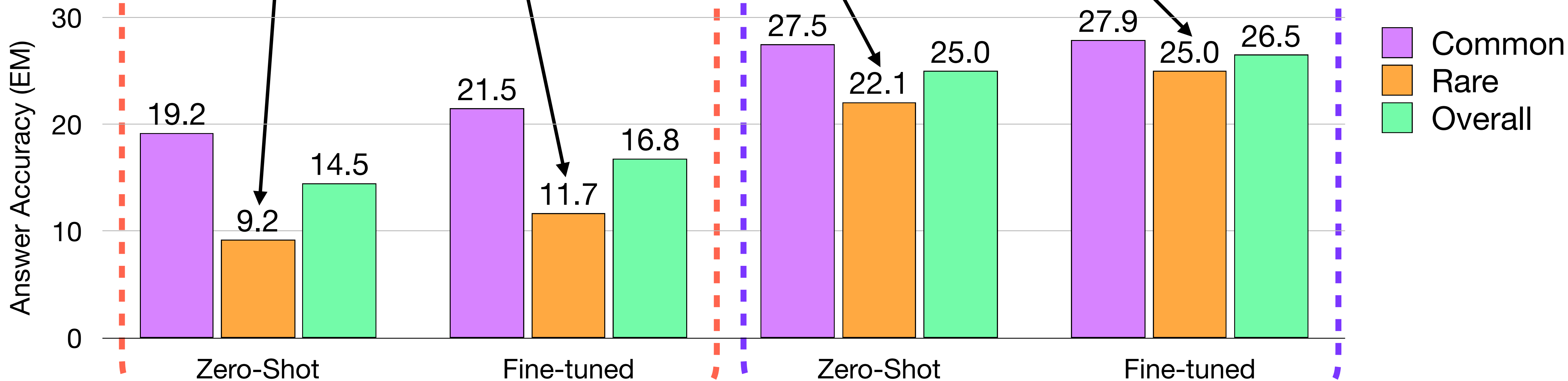
Closed-Book (BART)



Retrieval-Based (DPR)



1. Models perform worse on rare locations



Results: Geographical SituatedQA

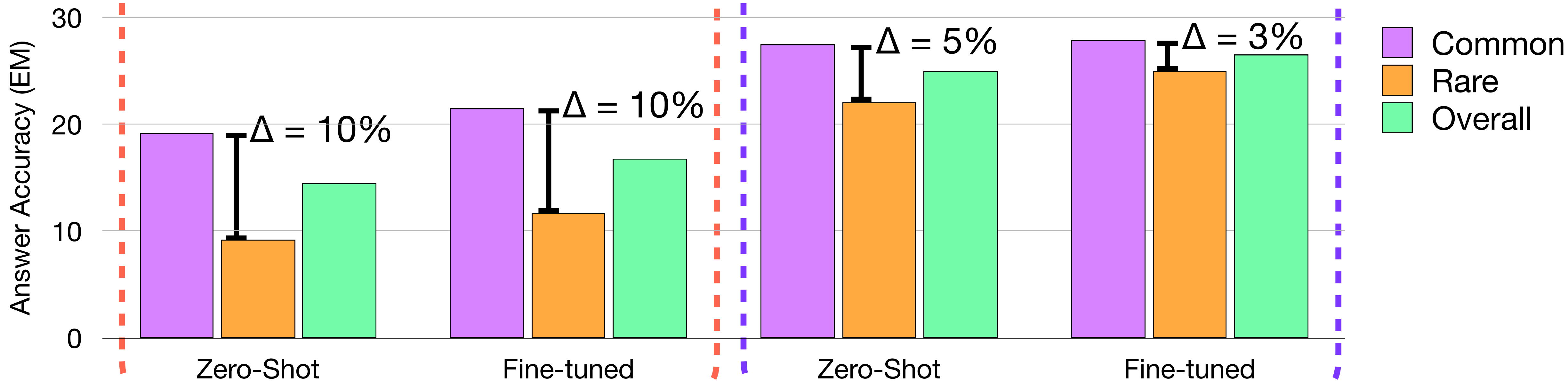
Closed-Book (BART)



Retrieval-Based (DPR)



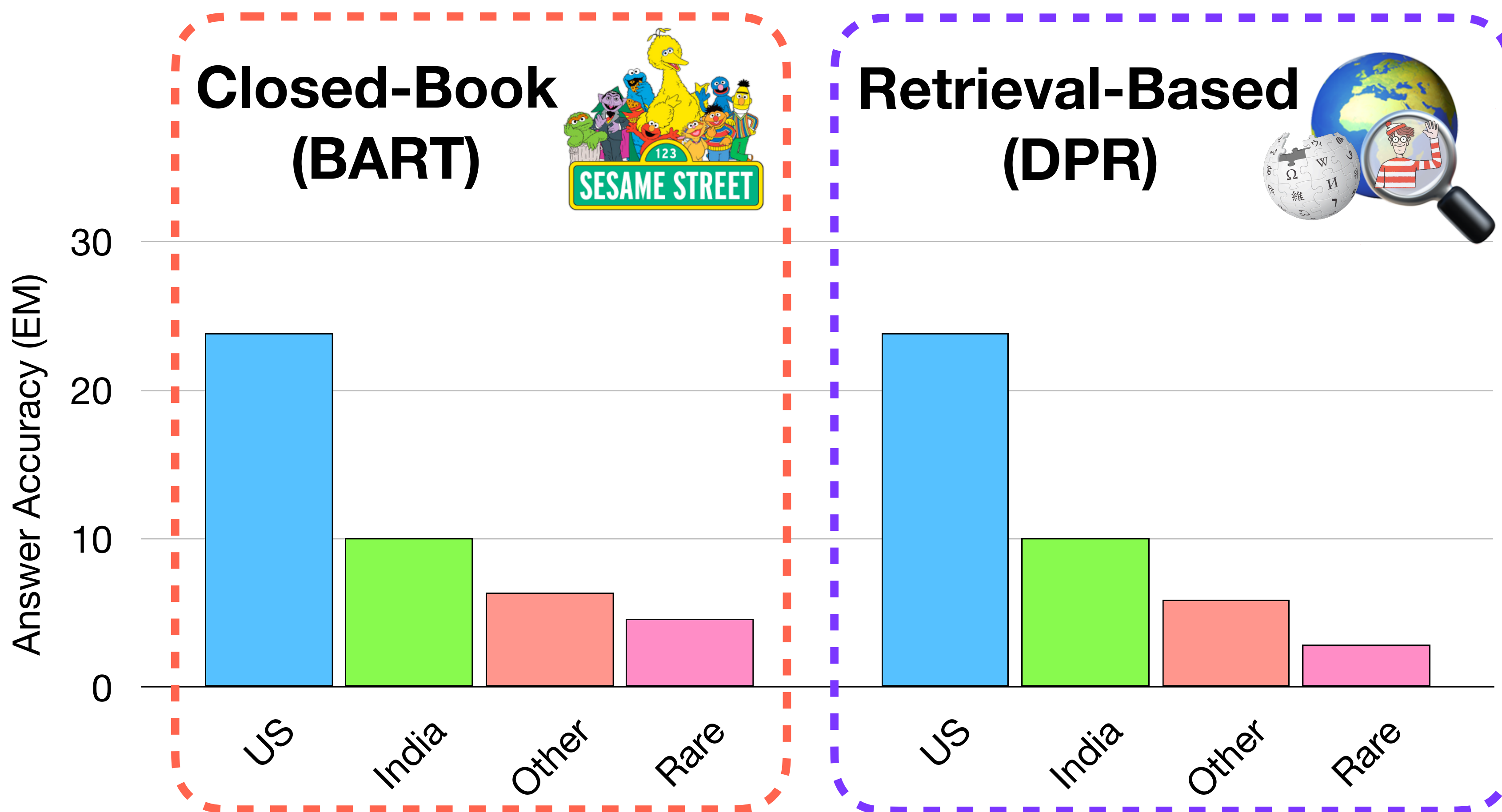
1. Models perform worse on rare locations
2. The performance gap between rare and common locations is less significant for retrieval-based methods



Analysis: Geographical Bias

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

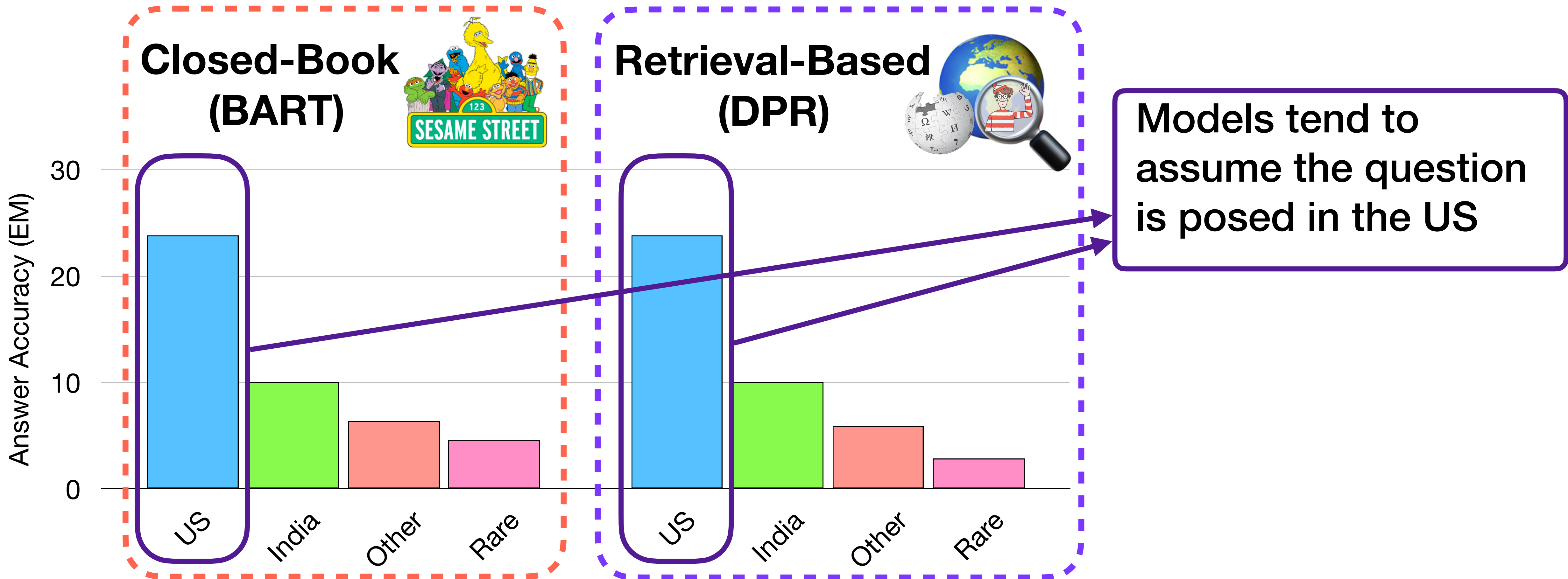
- We measure zero-shot performance on questions *without* the location
- We separate the **US and India**, the two most frequent locations, from the remaining common locations



Analysis: Geographical Bias

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

- We measure zero-shot performance on questions *without* the location
- We separate the **US and India**, the two most frequent locations, from the remaining common locations



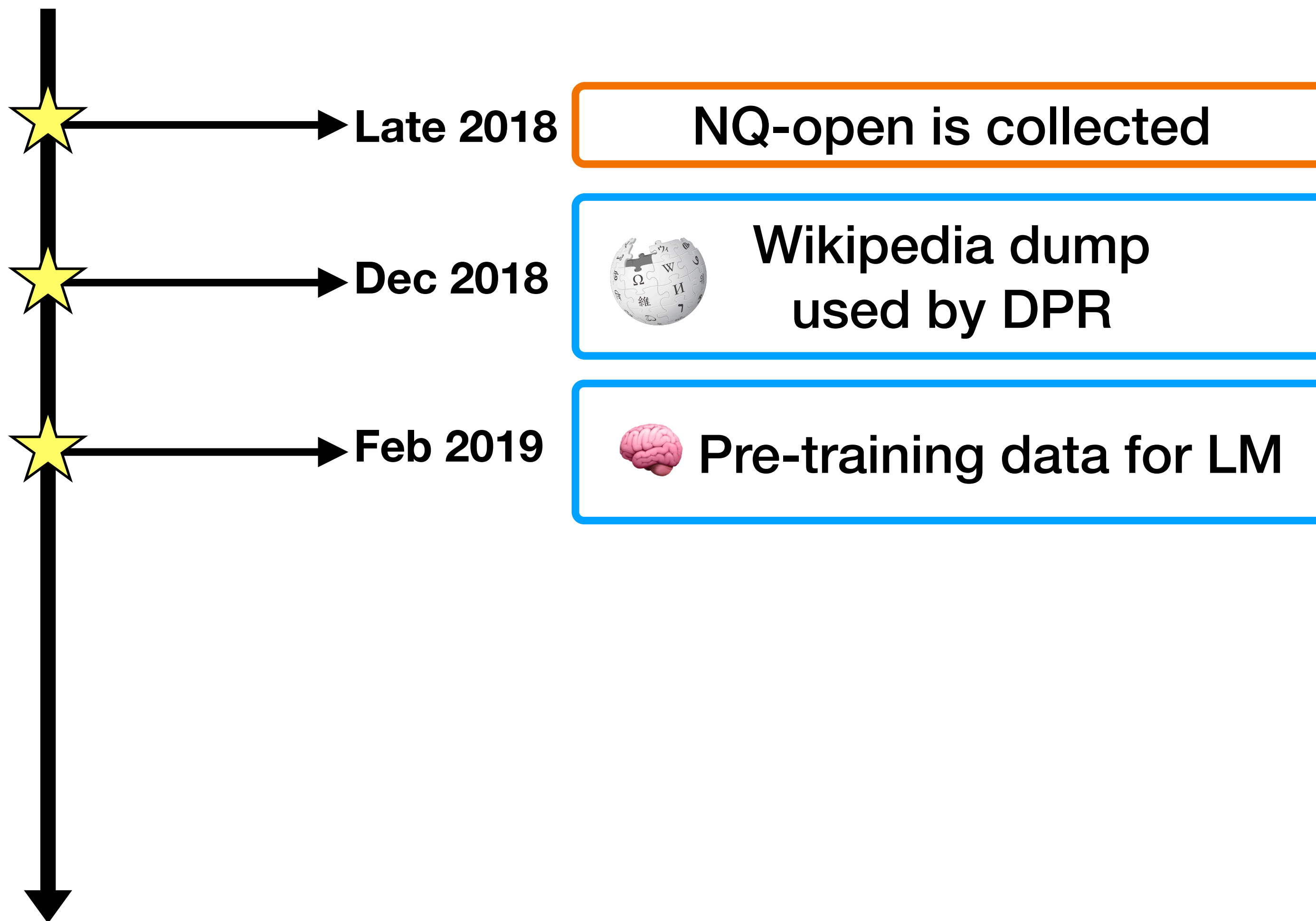
Analysis: Temporal Adaptation

Question

Do open-retrieval QA models that are trained on data collected in the past generalize to answering questions asked in the present?

Temporal Dependence of Models

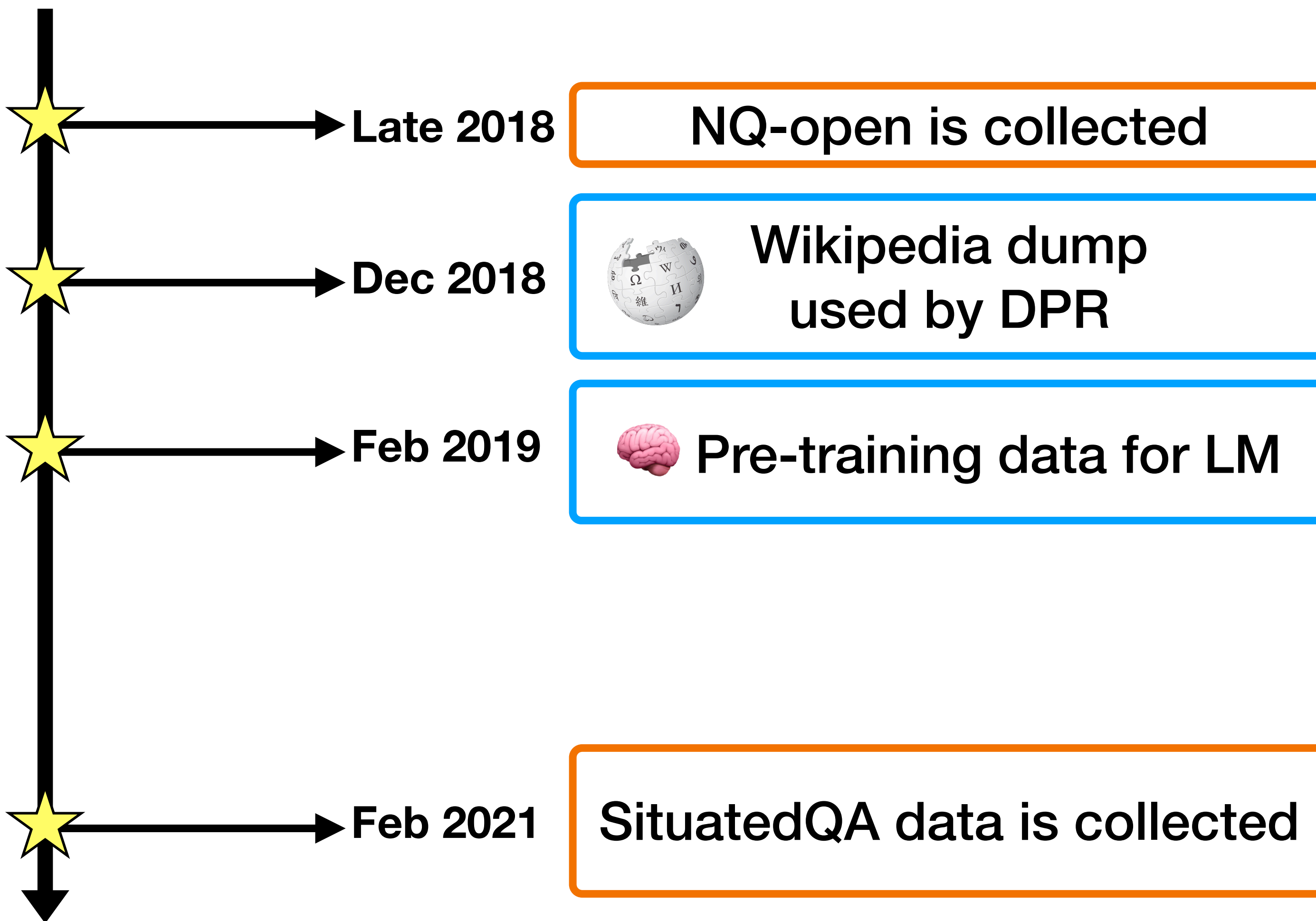
SOTA QA models resolve context by using data from the same time period



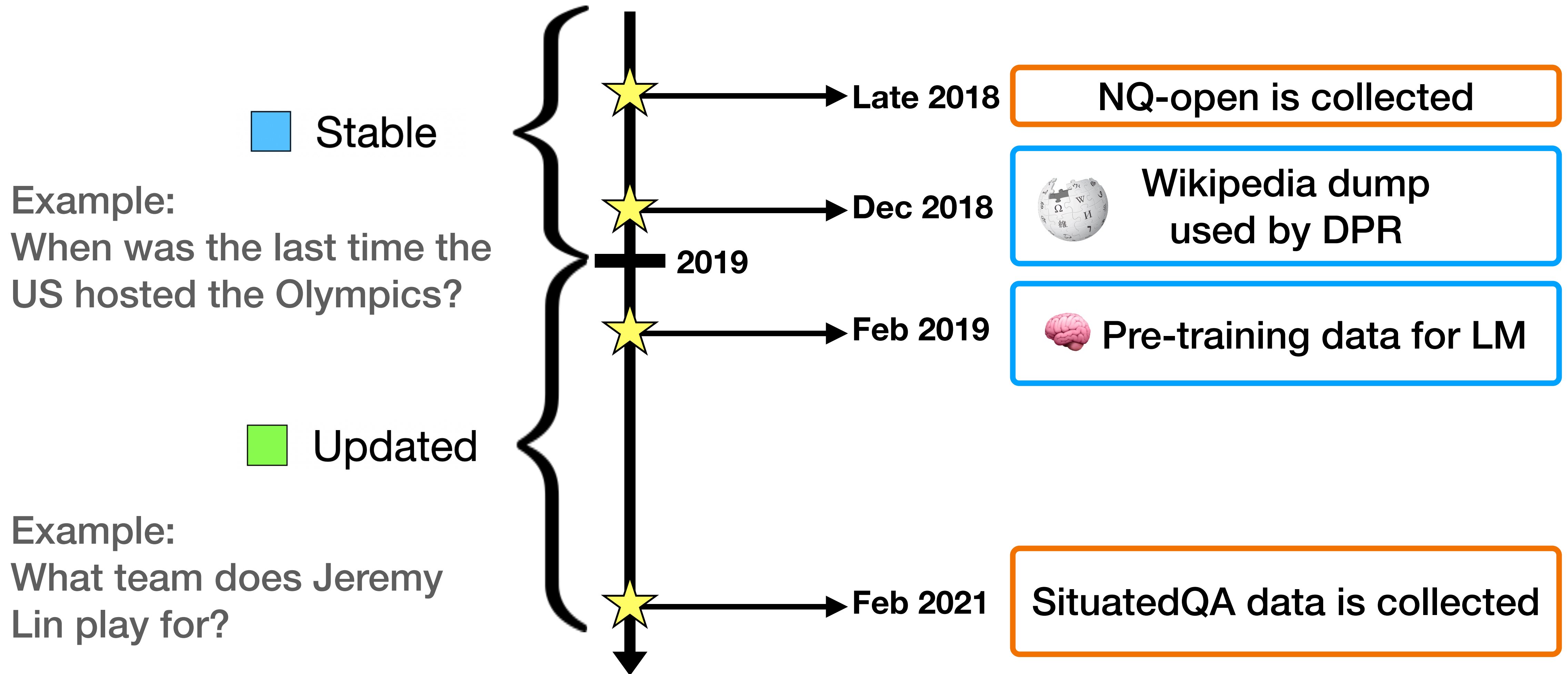
Temporal Dependence of Models

SOTA QA models resolve context by using data from the same time period

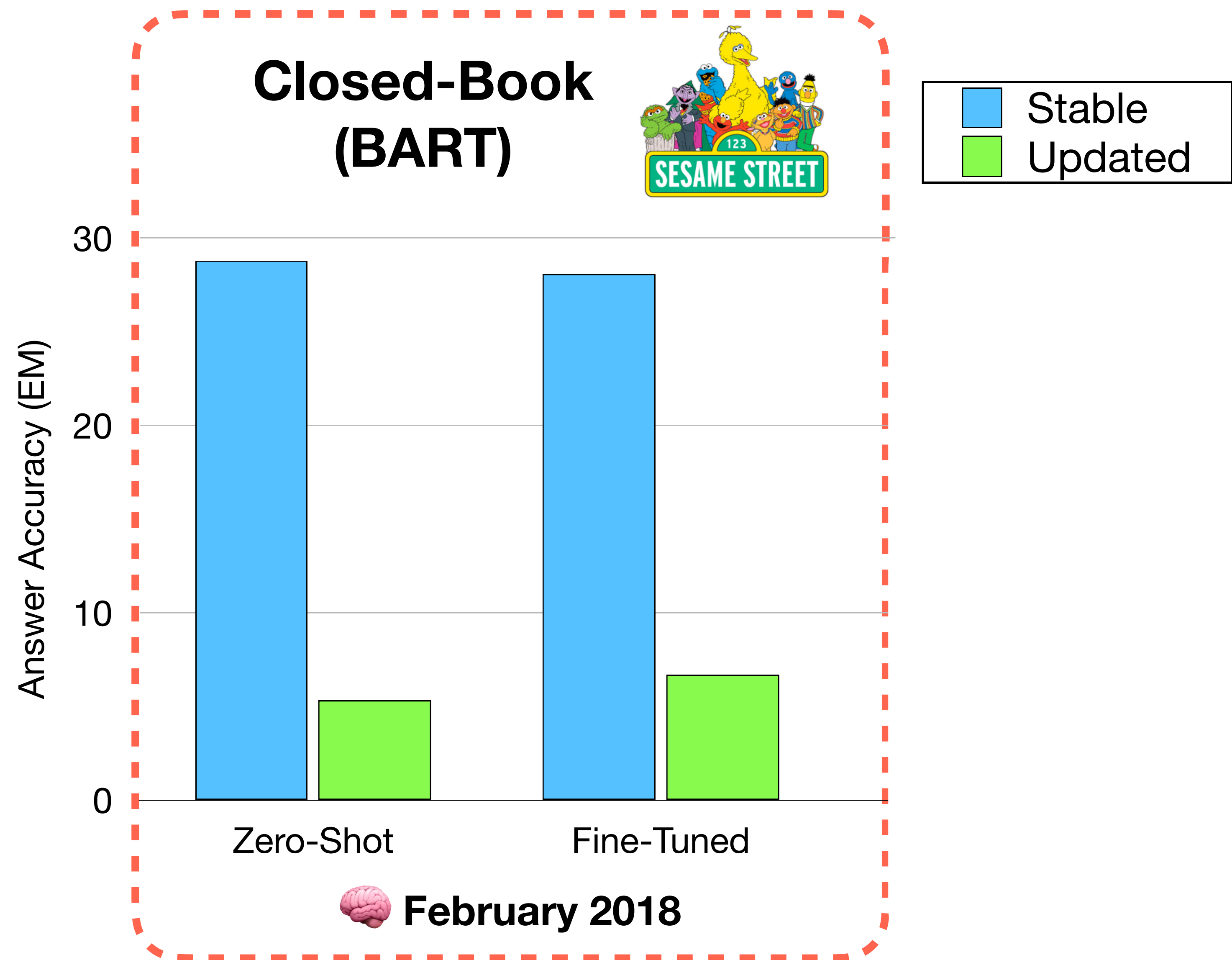
SituatedQA breaks this consistency



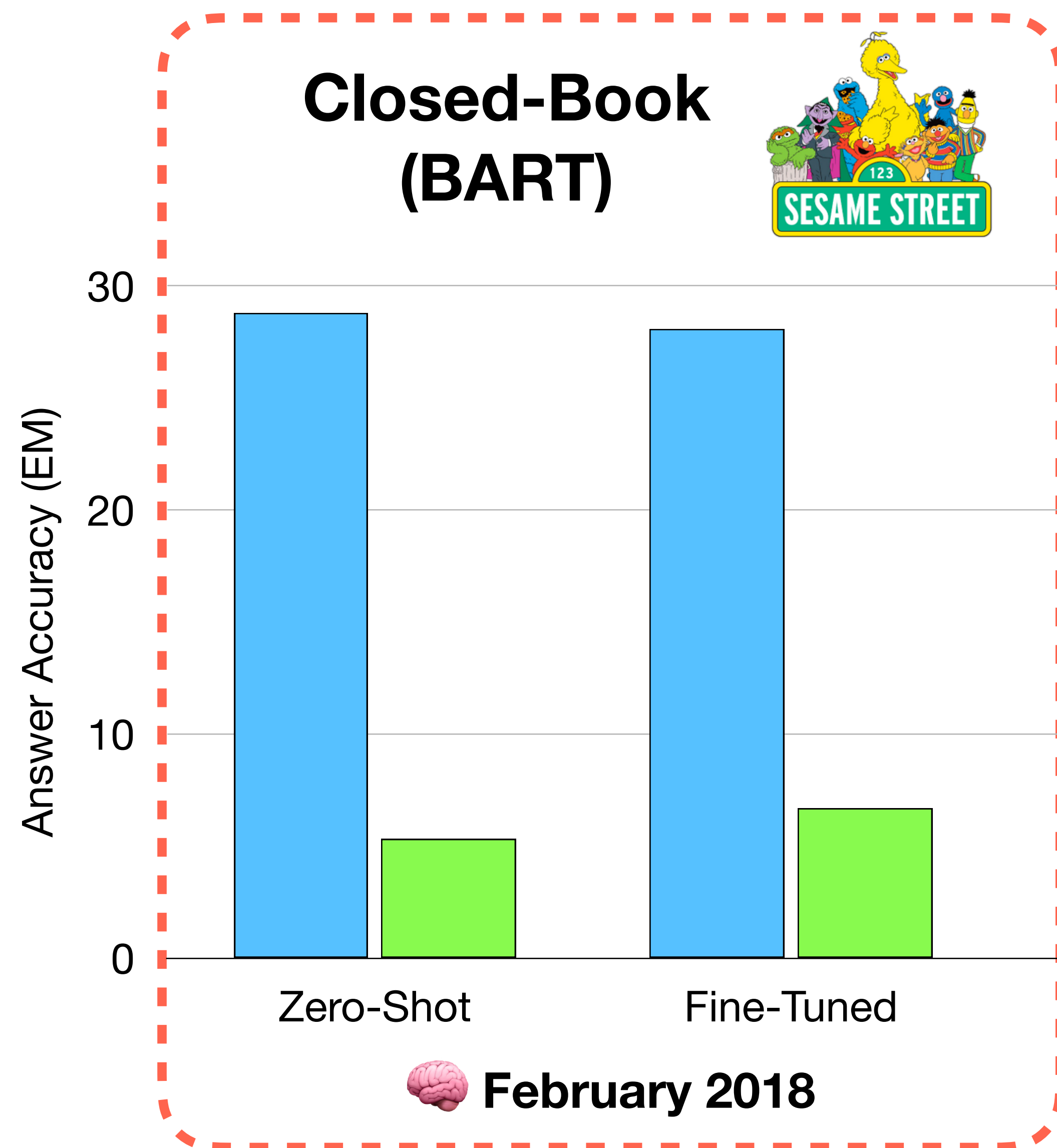
Temporal Dependence of Models



Analysis: Temporal Adaptation

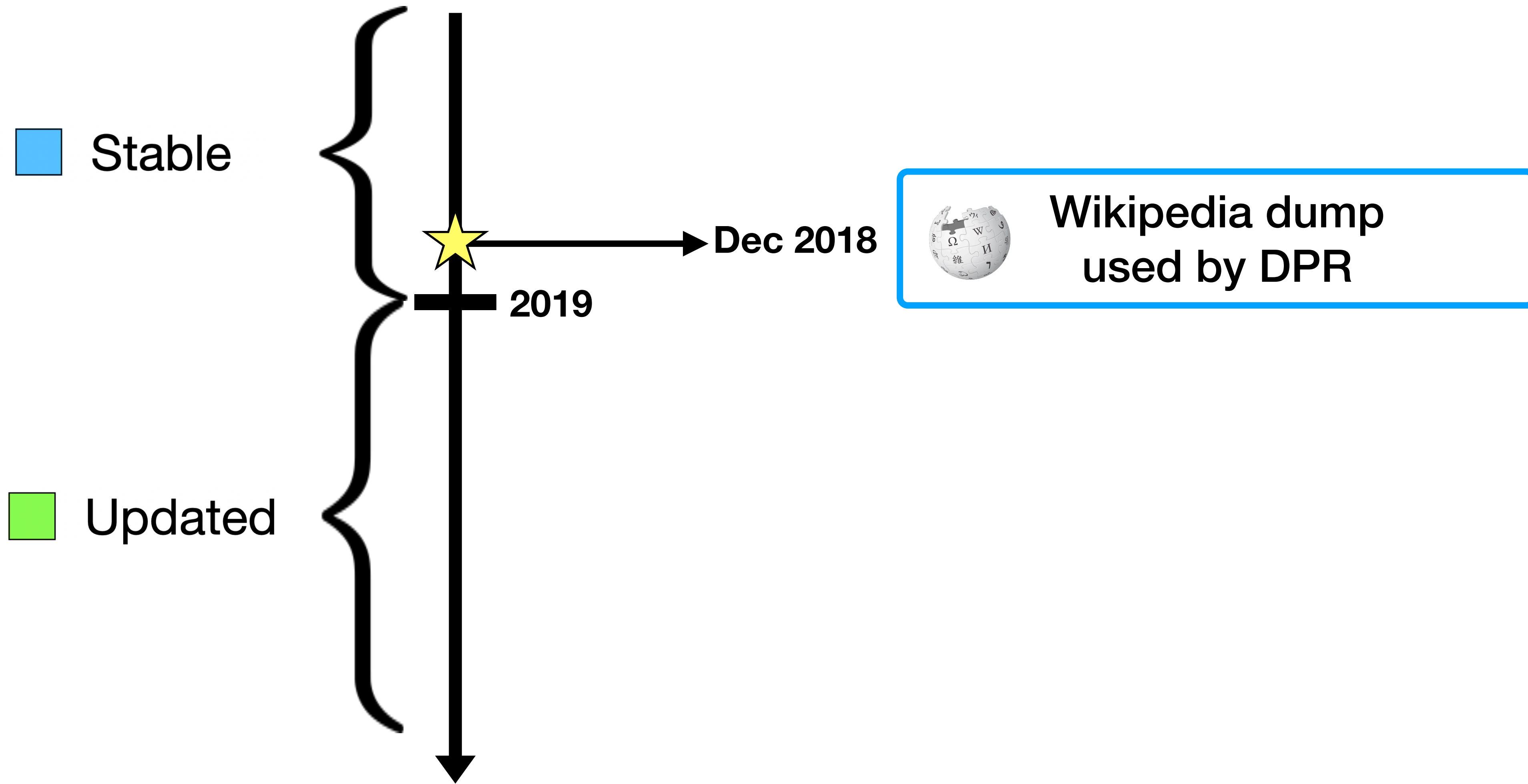


Analysis: Temporal Adaptation

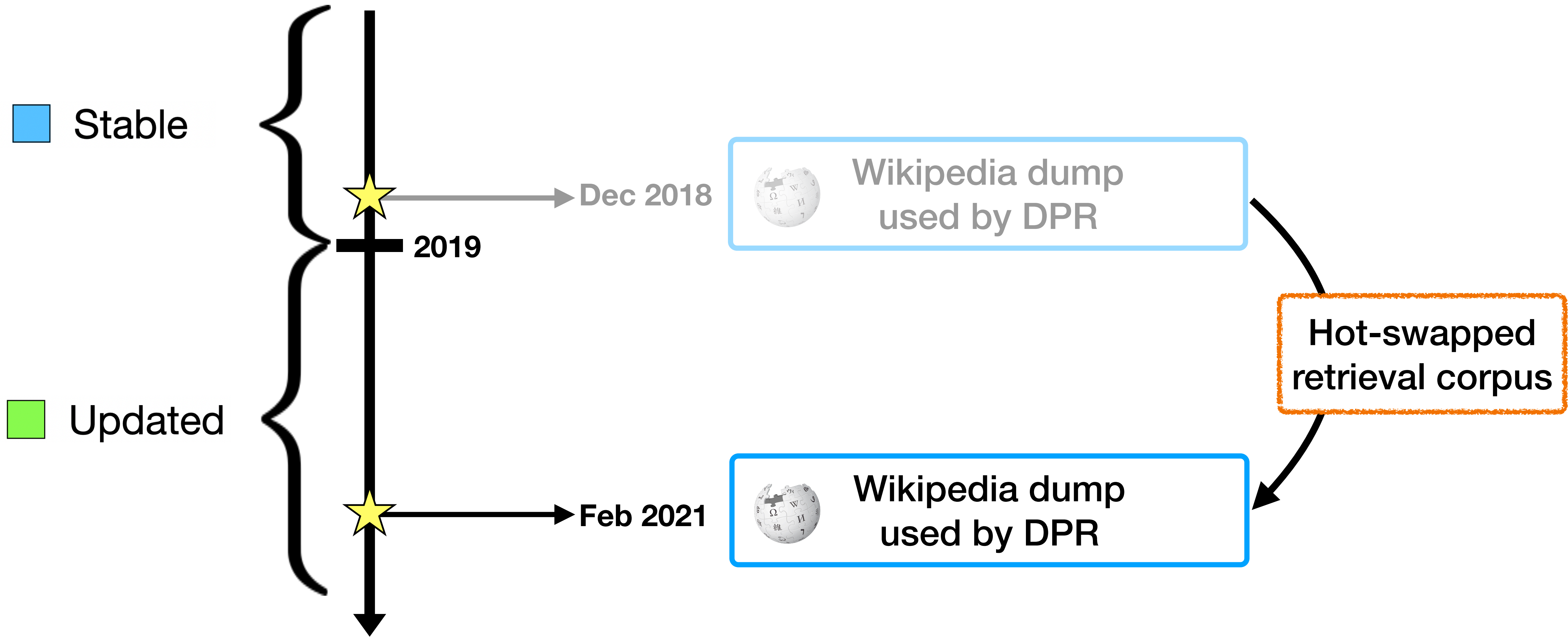


**Closed-book models
are unable to adapt
to questions with updated
answers**

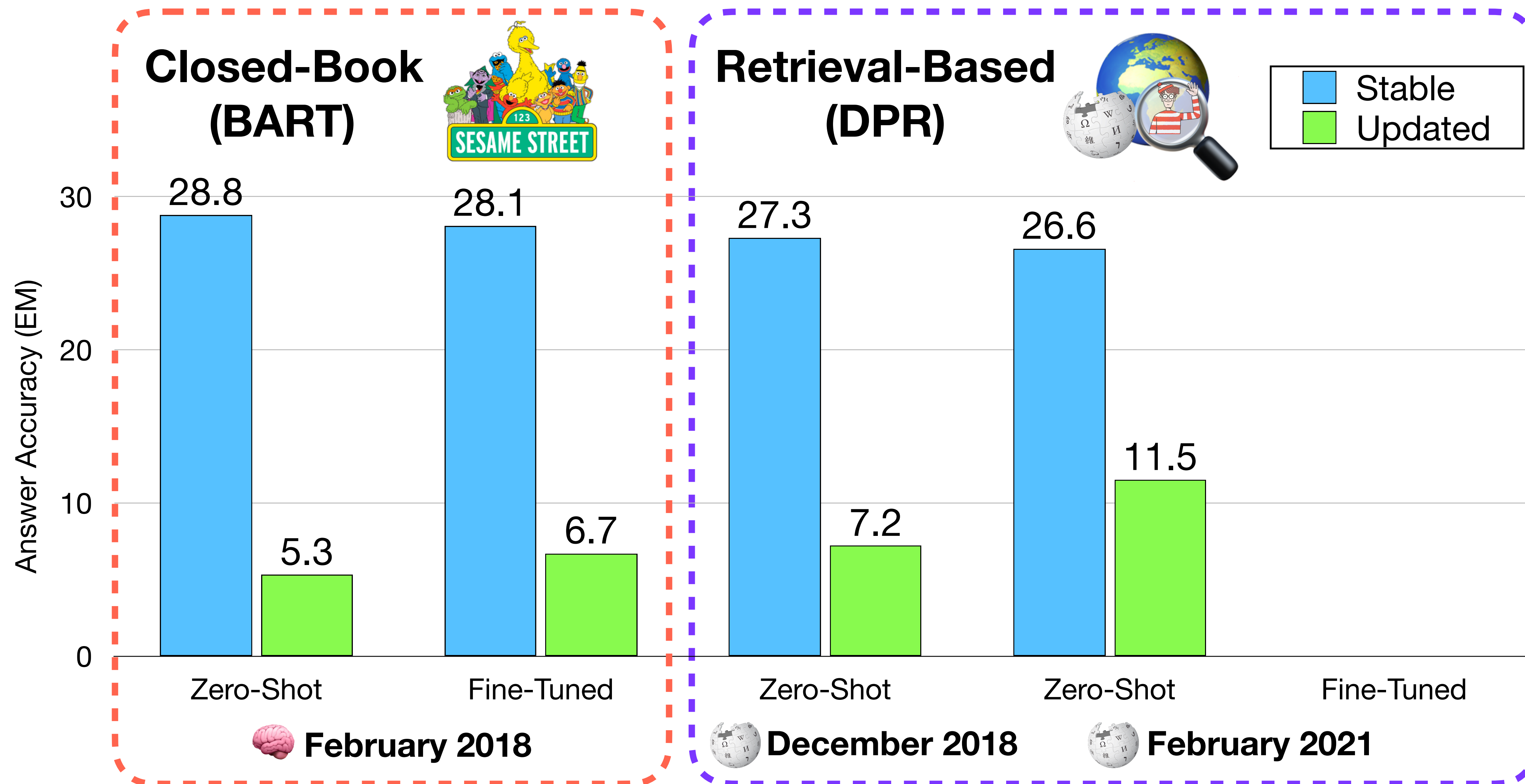
Temporal Adaptation of Retrieval-based Systems



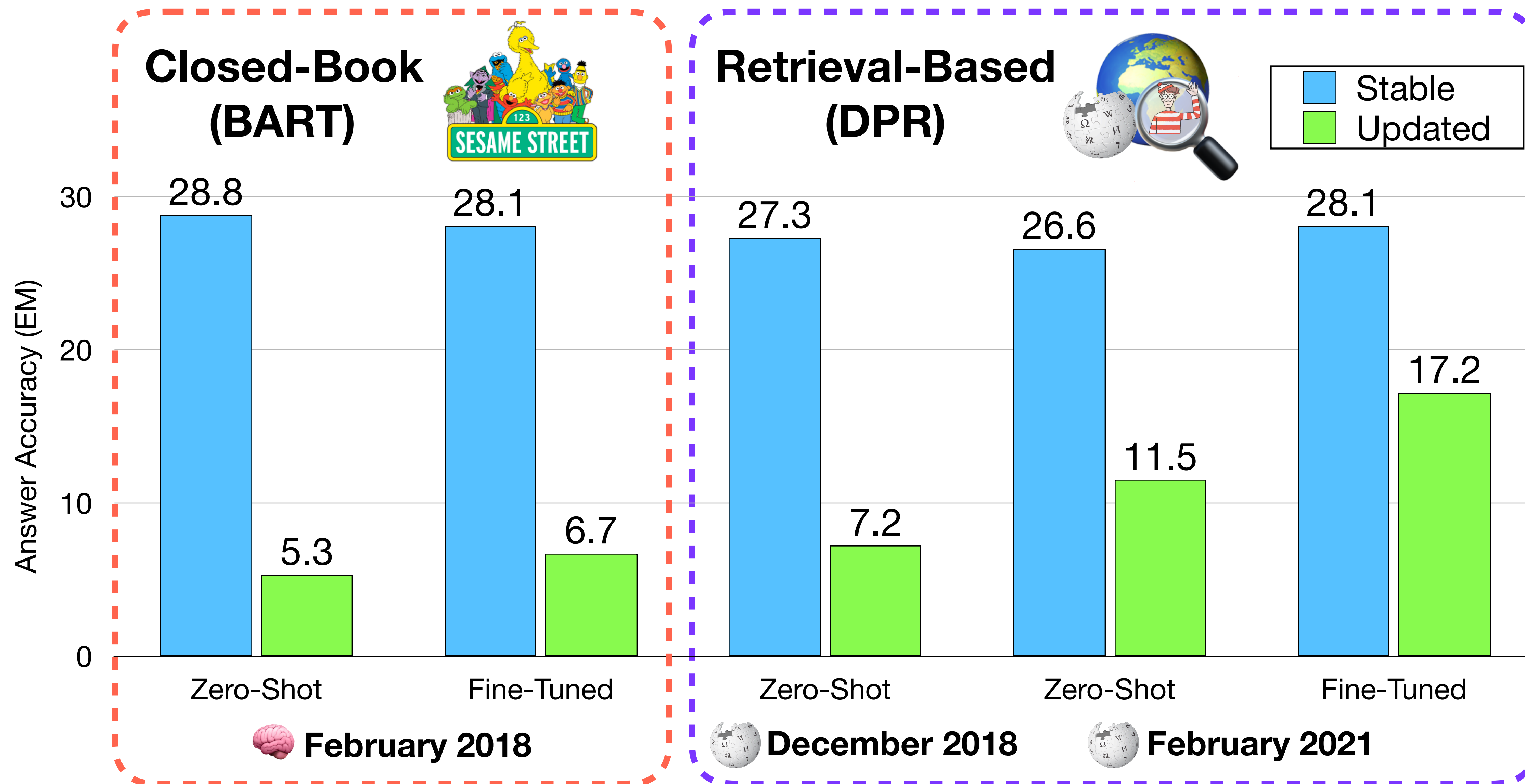
Temporal Adaptation of Retrieval-based Systems



Analysis: Temporal Adaptation



Analysis: Temporal Adaptation

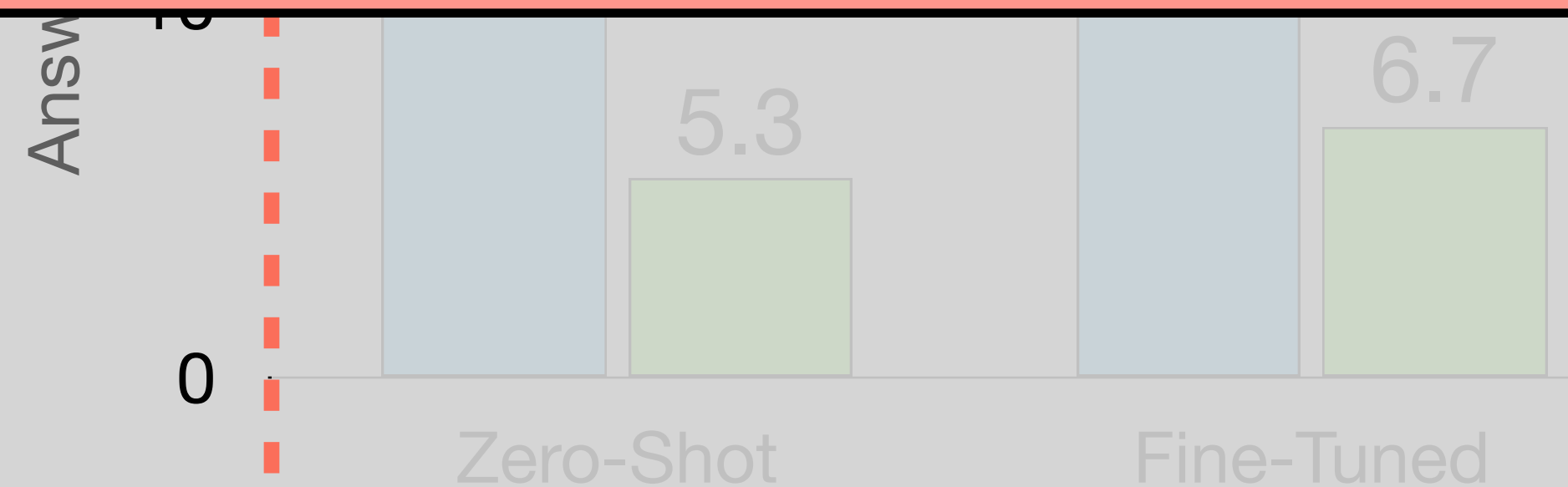


Analysis: Temporal Adaptation

Closed-Book
(BART)



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

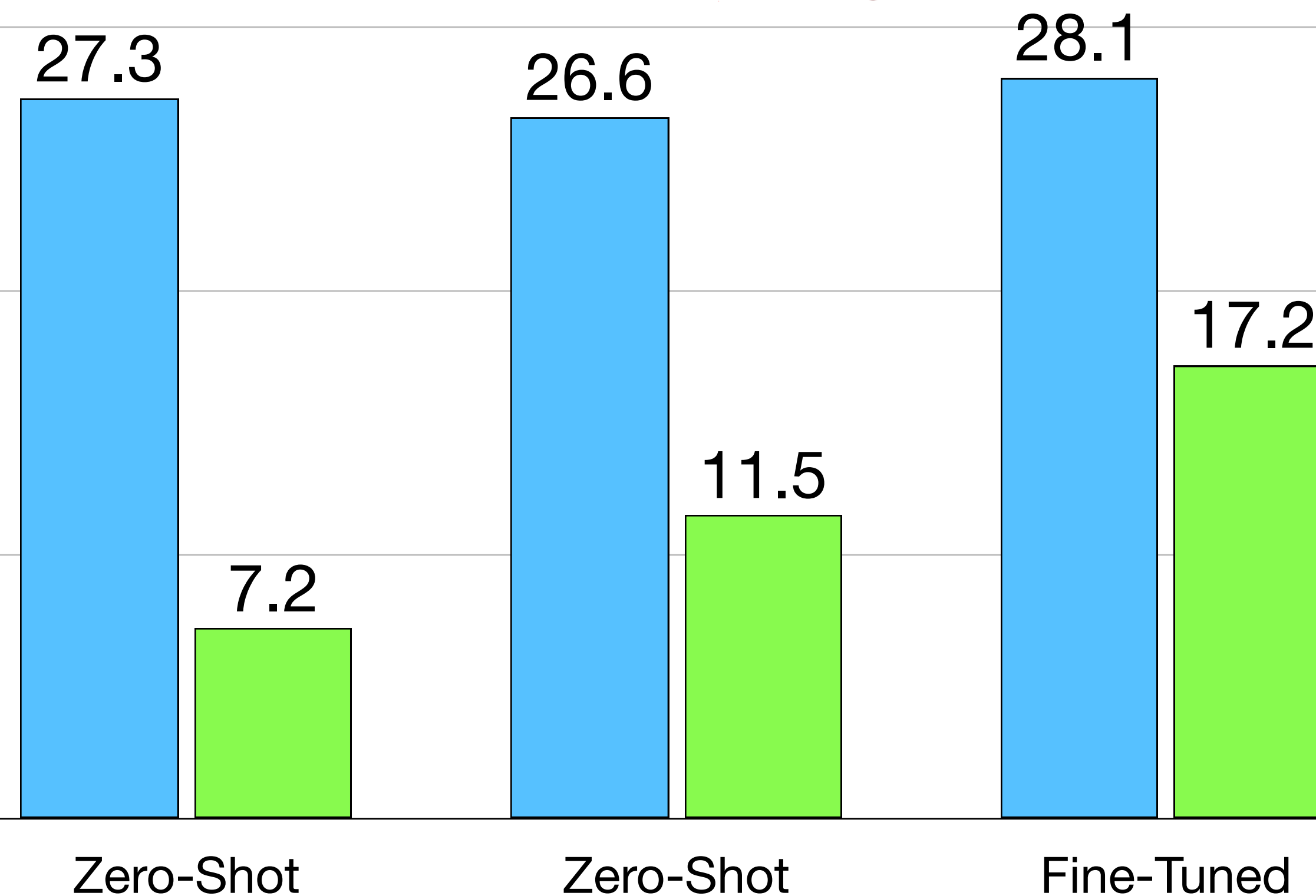


February 2018

Retrieval-Based
(DPR)



Stable
Updated



December 2018

February 2021

Related Work:

AmbigQA explores ambiguity in QA, including ambiguity that cannot be resolved with context

- Temporal-dependence is included under their definition of ambiguity

[Min et al., EMNLP 2020]

CronQuestions explores temporal dependence in knowledge-graph QA

- Questions are generated from temporally-dependent relations within WikiData

[Saxena et al., ACL 2021]

TempLama explores creating temporally-aware pretrained language models

- The authors train a version of T5 that utilizes temporal context (the year) during pretraining

[Dhingra et al., ArXiv 2021]

GD-VCR looks at geographical dependence in visual commonsense reasoning

- This dataset explores forms of cultural and location-specific commonsense in Visual QA

[Yin et al., EMNLP 2021]

Thank You!

Data & code are available now!

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 struggle with producing answers that are frequently updated on

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
--	--

Timeline: 12/2020, 01/2021, 02/2021, 03/2021, 04/2021, 05/2021

Context: Dec 18, 2014 Answer: Moderna, Pfizer	Context: Apr 10, 2021 Answer: Moderna, Pfizer, J&J
--	---

[SituatedQA.github.io](https://situatedqa.github.io)