Simple Entity-Centric Questions Challenge Dense Retrievers

Christopher Sciavolino*, Zexuan Zhong*, Jinhyuk Lee, Danqi Chen

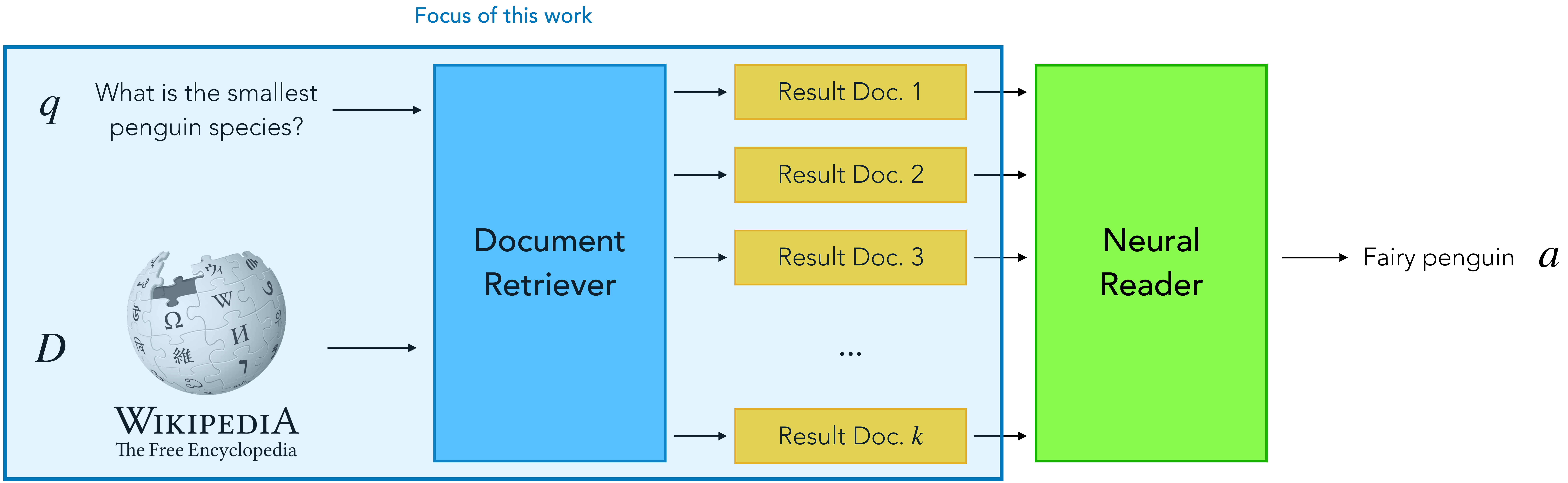
* Equal Contribution // EMNLP 2021

PRINCETON
NLP

Open-domain QA: Retriever-Reader System

1. Retrieve relevant documents

2. Read the documents and return an answer



Typical Retriever-Reader System

(Chen et al., 2017, Yang et al., 2019, Lee et al., 2019)

Retrieval: Sparse Models

Early systems used **BM25** for retrieval, which is very good at **lexical matching**

Lexical Overlap ✓

Question: What is the smallest penguin species?
Answer: Fairy penguin
Gold Passage: … the smallest penguin species is the **fairy penguin**, commonly found in …

Semantic Overlap ✗

Question: who is the bad guy in lord of the rings
Answer: Sauron
Gold Passage: … story's main antagonist, the Dark Lord **Sauron**, who in an earlier…

(Robertson & Zaragoza, 2009)

Retrieval: Dense Models

Newer systems use **dense retrieval**, which is very good at **semantic understanding**

Lexical Overlap

Question: What is the smallest penguin species?
Answer: Fairy penguin
Gold Passage: … the smallest penguin species is the **fairy penguin**, commonly found in …

Semantic Overlap

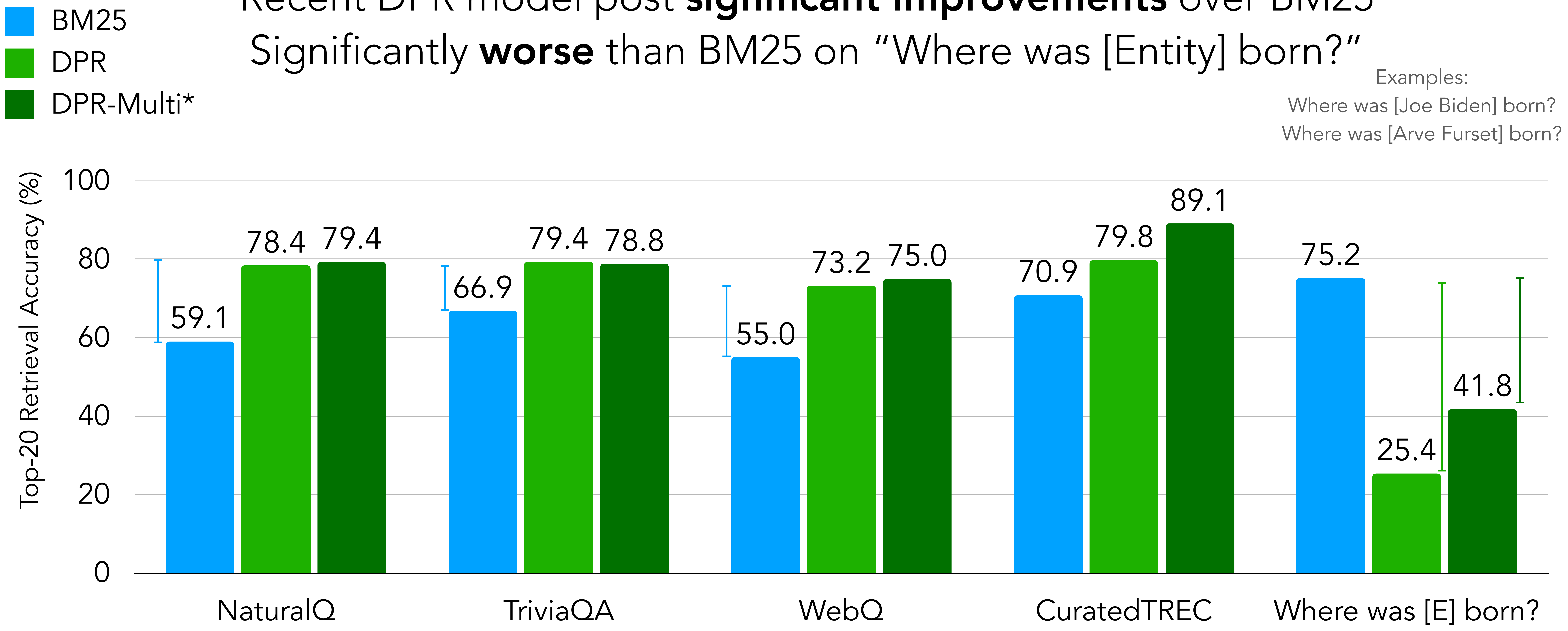
Question: who is the bad guy in lord of the rings
Answer: Sauron
Gold Passage: … story's main antagonist, the Dark Lord **Sauron**, who in an earlier…

(Lee et al., 2019, Karpukhin et al., 2020, Guu et al., 2020)

Retrieval: Successful Dense Models

Recent DPR model post **significant improvements** over BM25

Significantly **worse** than BM25 on "Where was [Entity] born?"



* Multi trained on NQ (Kwiatkowski et al., 2019), TriviaQA (Joshi et al., 2017), WebQ (Berant et al., 2013), and CuratedTREC (Baudis and Sedivy, 2015)

EntityQuestions: Dataset Overview

Sampled 24 relations and created hand-crafted question templates

Key idea: Simple questions about specific entities

Arve Furset

From Wikipedia, the free encyclopedia

Arve Eilif Furset (born 5 December 1964 in Askvoll, Western Norway) is a Norwegian composer, jazz musician (piano, keyboards) and music producer, known from a series of record releases and cooperations with

Q: Where was **Arve Furset** born?
A: Askvoll

Joe Biden

From Wikipedia, the free encyclopedia

"Joseph Biden" and "Biden" redirect here. For his tenure as president, s Joseph Biden III, see Beau Biden. For other uses, see Biden (disambig

Joseph Robinette Biden Jr.[a] (/ˈbaɪdən/ *BY-dən*; born November 20, 1942) is an American politician who is the 46th and current president of the United States. A member of the Democratic Party, he served as the 47th vice president from 2009 to 2017 under Barack Obama and represented Delaware in the United States Senate from 1973 to 2009.

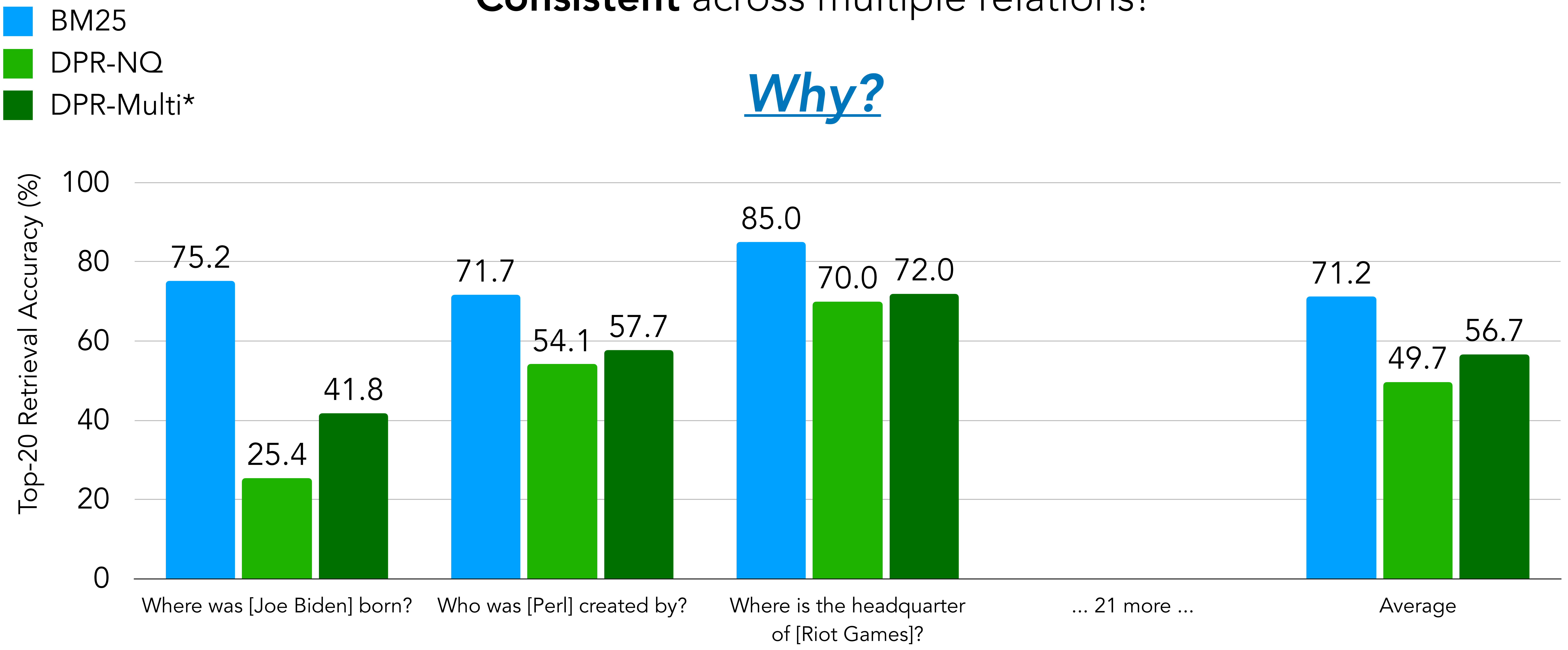
Born and raised in Scranton, Pennsylvania, and later in New Castle County, Delaware, Biden studied at the University of Delaware before earning his law degree from Syracuse University in 1968. He was elected

Q: Where was **Joe Biden** born?
A: Scranton

EntityQuestions: Results

Consistent across multiple relations!

Why?



* Multi trained on NQ (Kwiatkowski et al., 2019), TriviaQA (Joshi et al., 2017), WebQ (Berant et al., 2013), and CuratedTREC (Baudis and Sedivy, 2015)

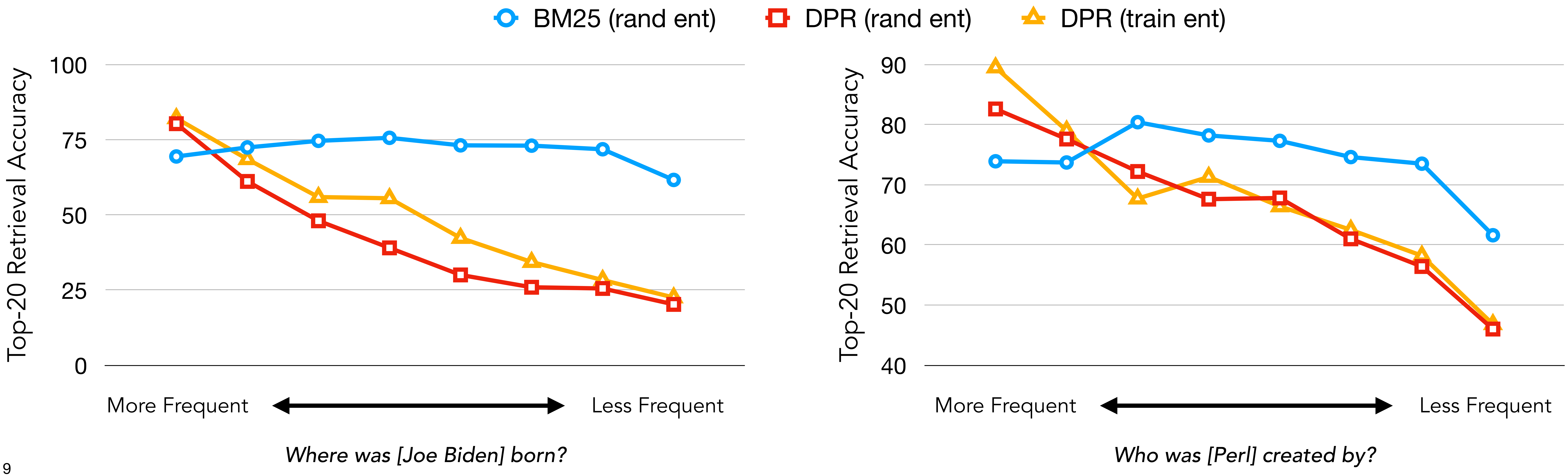
Understanding the Problem: Two Questions

1. How does the specific entity affect performance?

2. Can the model generalize to new entities if it sees the question pattern?

Understanding the Problem: Two Questions

1. How does the specific entity affect performance? Answer: DPR has a popularity bias!



Understanding the Problem: Two Questions

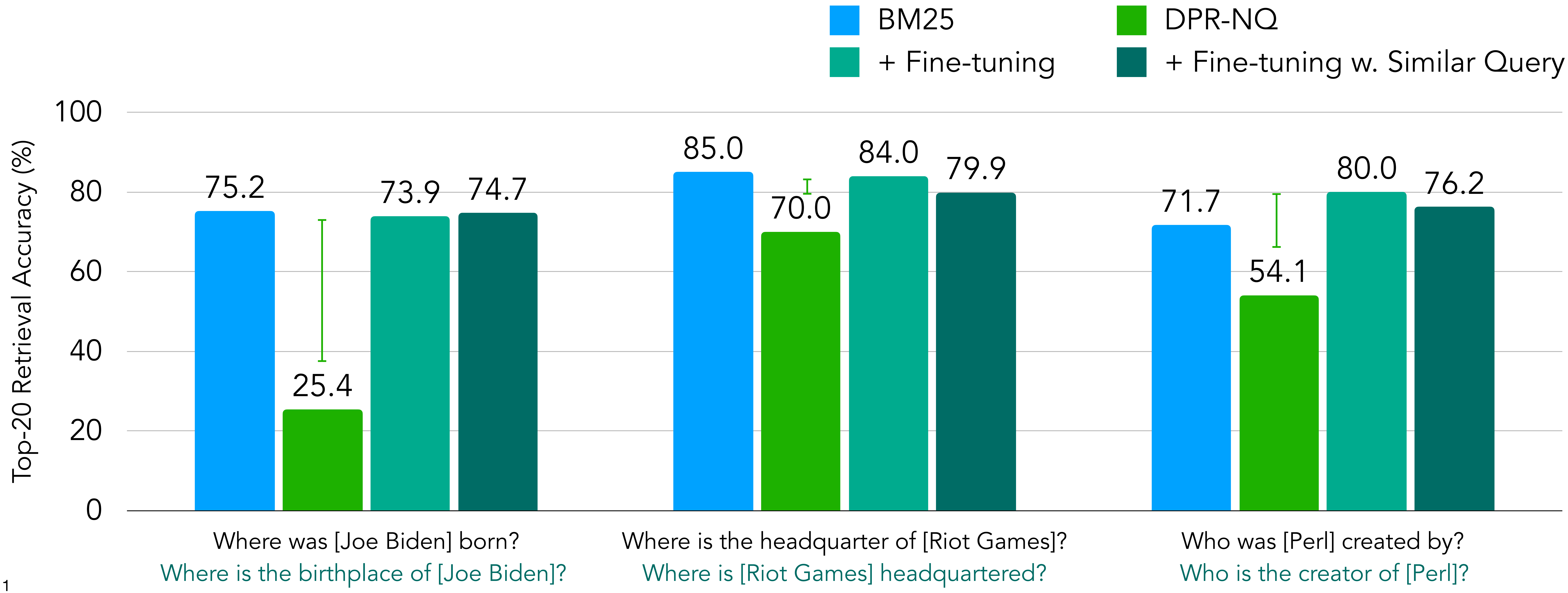
2. If the model observes the pattern during training, can it generalize to new entities?

Training: Where was [Joe Biden] born?

Testing: Where was [Arve Furset] born?

Understanding the Problem: Two Questions

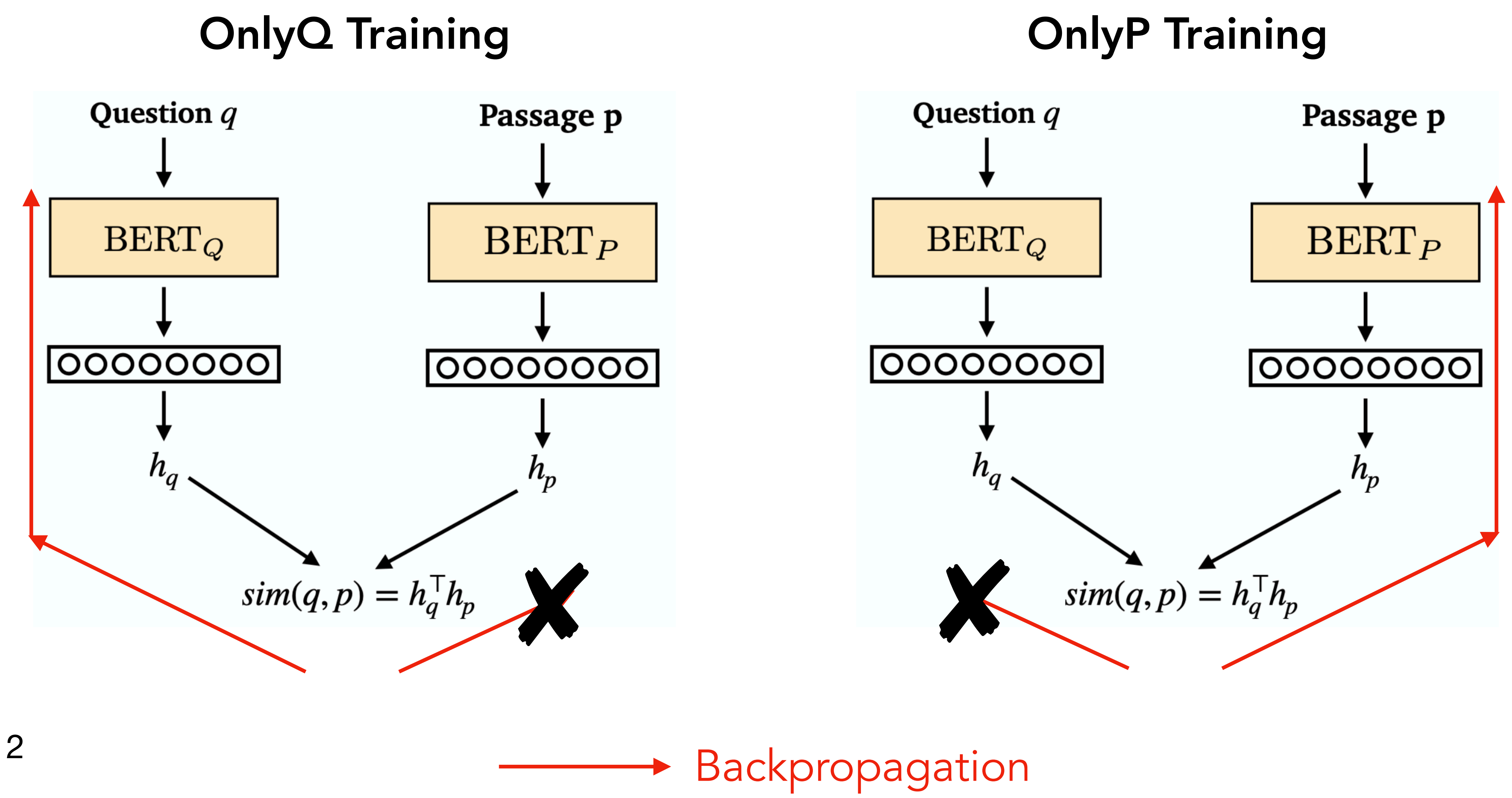
2. If the model observes the pattern during training, can it generalize to new entities? Answer: Yes, when fine-tuning both encoders!

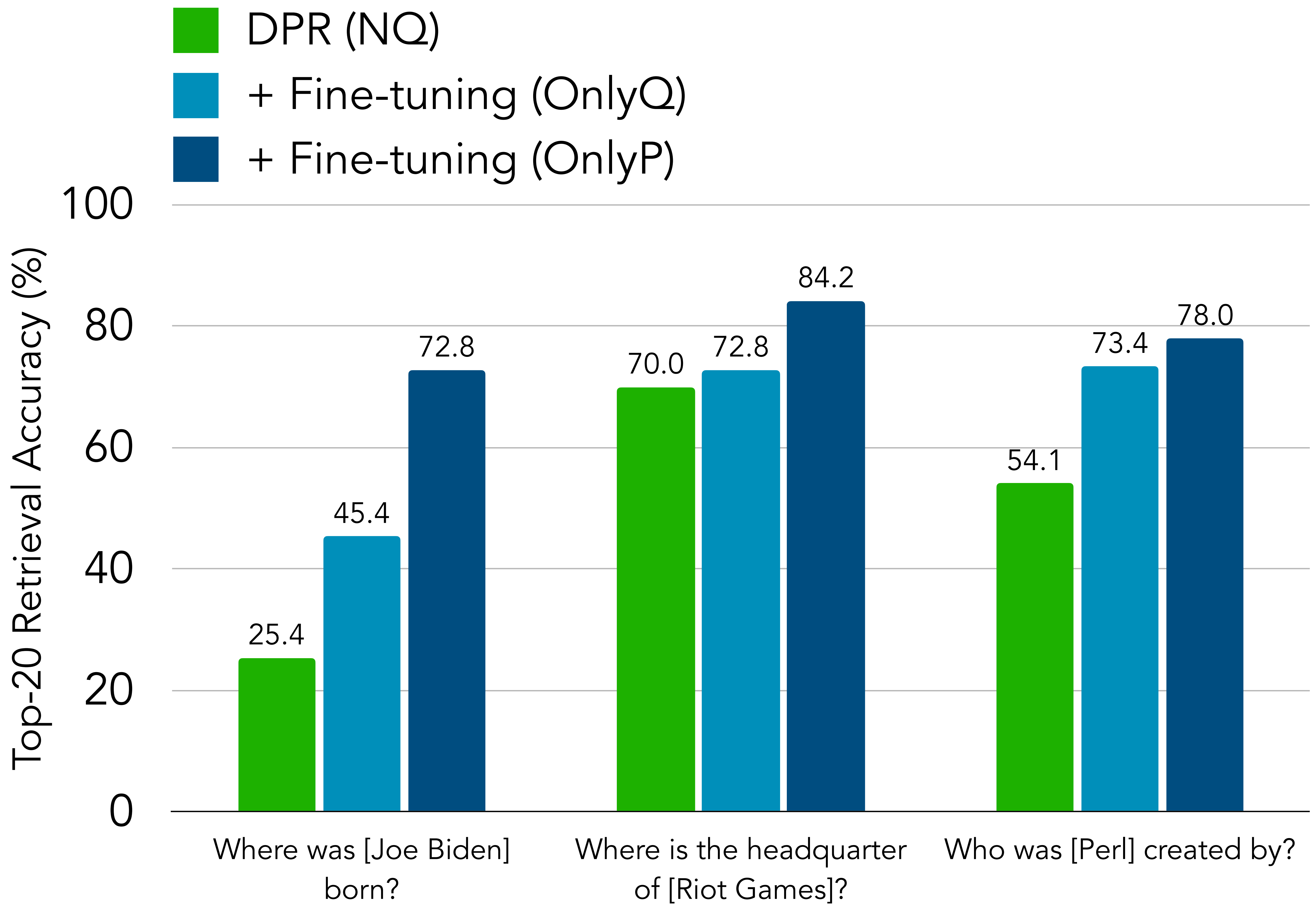


Understanding the Problem: Two Questions

2. If the model observes the pattern during training, can it generalize to new entities? Answer: Yes, when fine-tuning both encoders…

But the passage encoder is crucial!

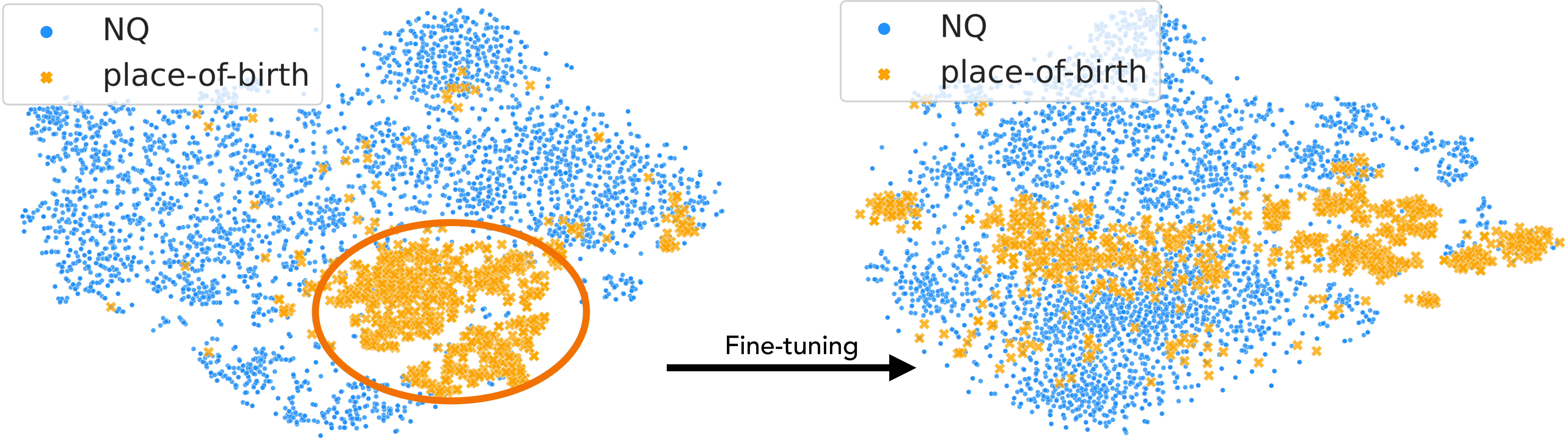




Understanding the Problem: Two Questions

2. If the model observes the pattern during training, can it generalize to new entities? Answer: Yes, when fine-tuning both encoders…

But the passage encoder is crucial!



DPR-NQ

+ place-of-birth fine-tuning

Exploring Solutions

1. Data augmentation

2. Specialized question encoders

Exploring Solutions: Data Augmentation

1. **Data augmentation**: ***Train on*** some EntityQ examples!

Gains on EntityQ relations… with degradation on NQ

	NaturalQ	In-domain Relation	EntityQ Avg.
DPR-NQ	80.1	25.4	49.7
+ FT p-of-birth	62.8	74.3	56.2
DPR-NQ	80.1	25.4	49.7
+ FT headquarter	71.6	80.3	53.3
DPR-NQ	80.1	25.4	49.7
+ FT creator	70.8	80.8	52.3
BM25	64.5	-	71.2

Exploring Solutions: Data Augmentation

1. **Data augmentation**: ***Add*** some EntityQ examples to our training set!

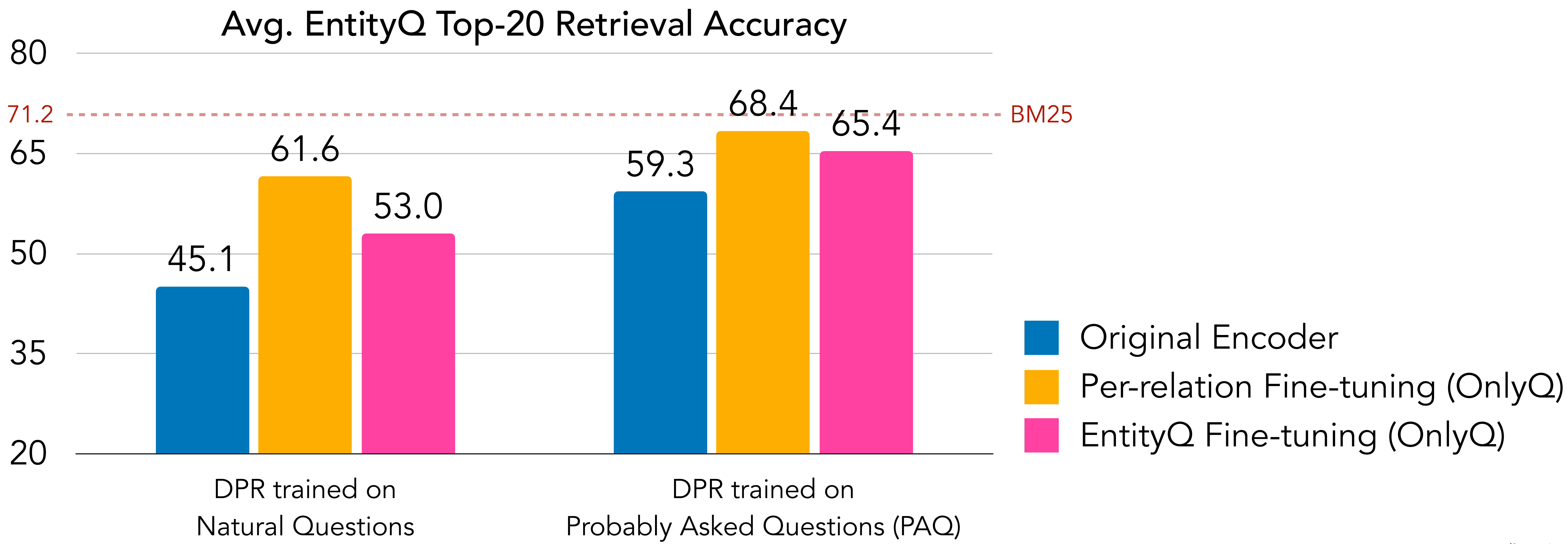
Reduced degradation on NQ… but muted improvements on EntityQ

	NaturalQ	**In-domain Relation**	**EntityQ Avg.**
DPR-NQ	80.1	25.4	49.7
+ FT p-of-birth	62.8	74.3	56.2
+ FT NQ ∪ p-of-birth	70.8	52.0	47.4
DPR-NQ	80.1	25.4	49.7
+ FT headquarter	71.6	80.3	53.3
+ FT NQ ∪ headquarter	75.1	81.3	49.5
DPR-NQ	80.1	25.4	49.7
+ FT creator	70.8	80.8	52.3
+ FT NQ ∪ creator	72.6	72.3	44.1
BM25	64.5	-	71.2

Exploring Solutions: Question Encoders

2. **Specialized question encoders**: Use a fixed passage index, and adapt question encoders

Very promising results! Choosing the right passage index matters.



(Lewis et al., 2021)

Conclusion

1. What's Wrong with Dense Retrievers?
   i. Dense models struggle on simple, entity-centric questions
2. Understanding the Problem
   i. Dense models exhibit popularity bias
   ii. The passage encoder is crucial for generalization
3. Exploring Solutions
   i. Data augmentation does not sufficiently solve the problem
   ii. Building a robust passage space is a very promising new direction

Future Work

1. Leveraging explicit entity memory during retrieval
2. Hybrid dense and sparse retrieval systems

Dataset & Code

https://github.com/princeton-nlp/EntityQuestions

Contact Us

cds3@cs.princeton.edu // zzhong@cs.princeton.edu