

# Phrase Retrieval Learns Passage Retrieval, Too

Princeton University 이진혁 연구원



# C CONTENTS

- 01 Background
- 02 Research Motivation
- 03 Formulation / Experiments #1, #2
- 04 Analysis / Experiments #3
- 05 Complexity Analysis
- 06 Conclusion



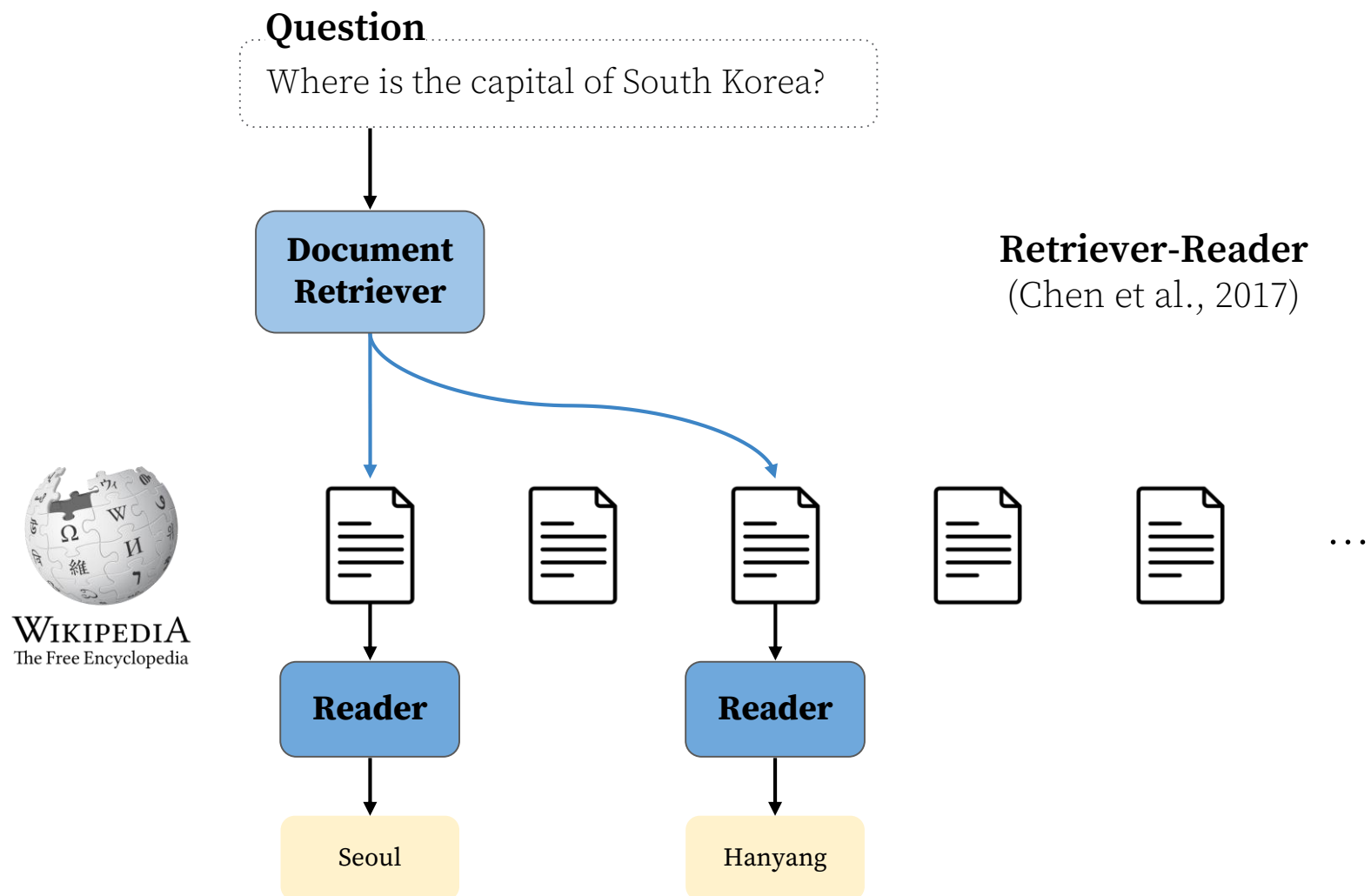


01

**Background**

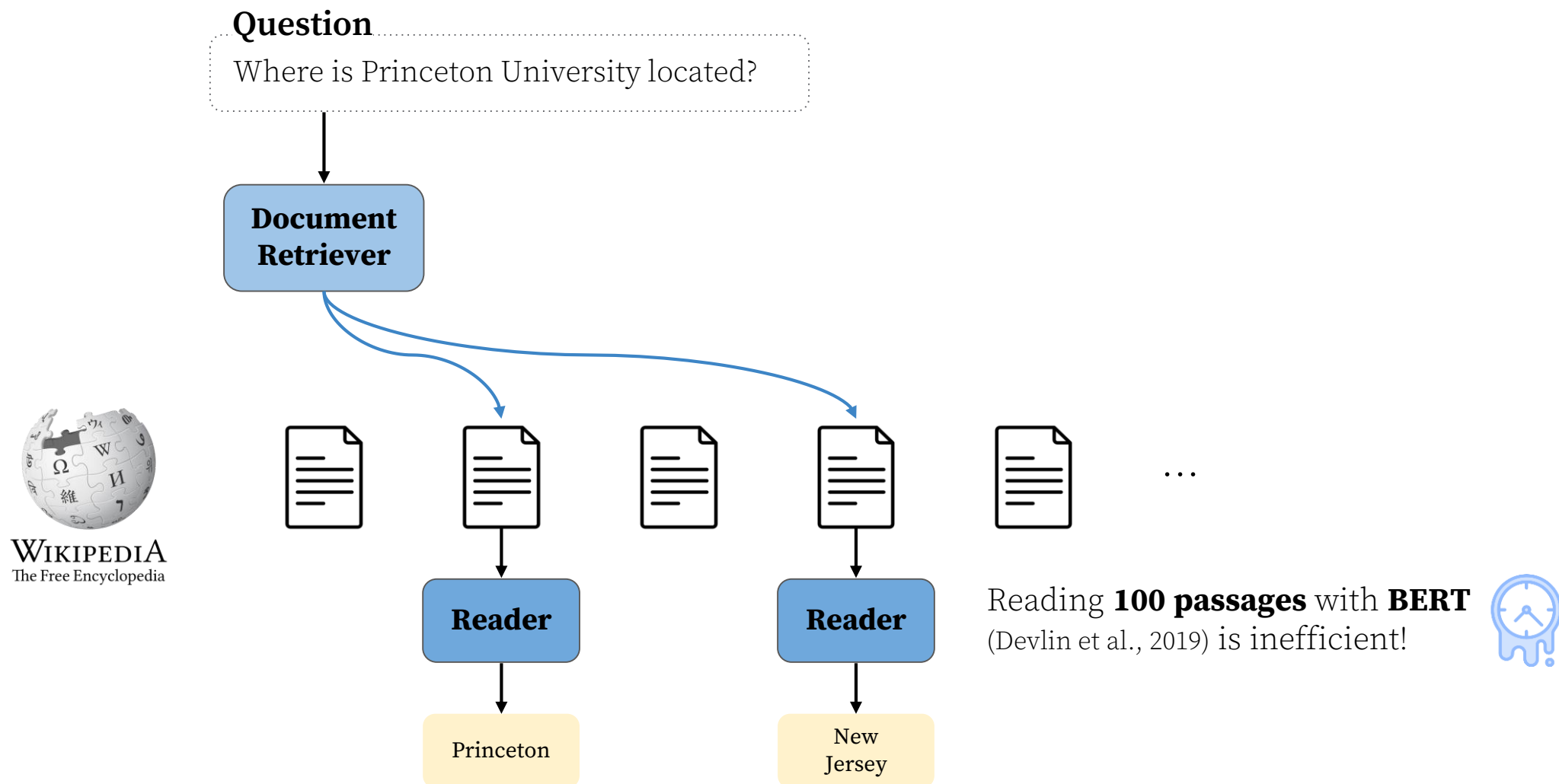
## 01

# Open-Domain Question Answering



# 01

## Open-Domain Question Answering



## 01

# Phrase Retrieval for Open-Domain QA

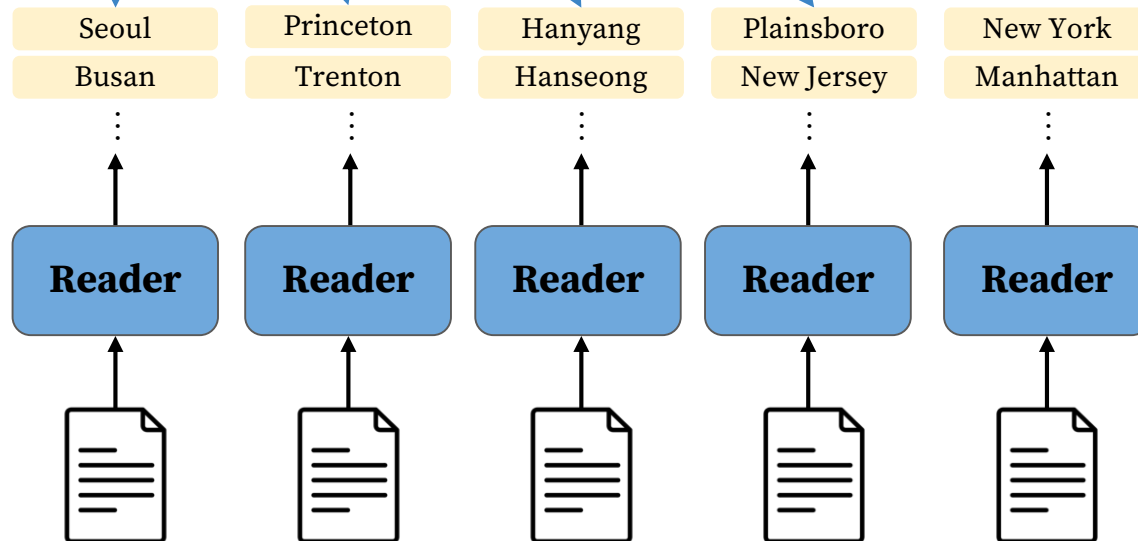
Phrase = any contiguous segment of text up to L words (Seo et al., 2019)

## Question

Where is Princeton University located?

## Phrase Retriever

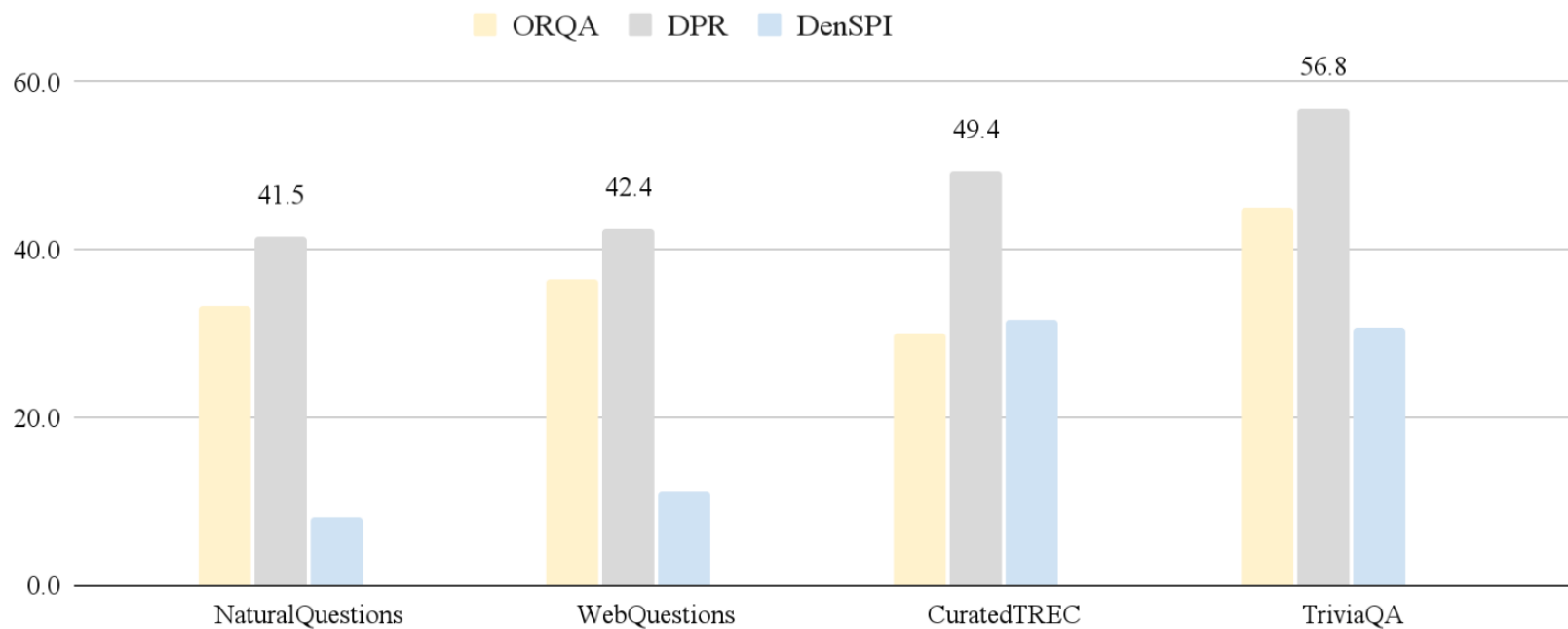
**DensePhrases** (Lee et al., 2021)



WIKIPEDIA  
The Free Encyclopedia

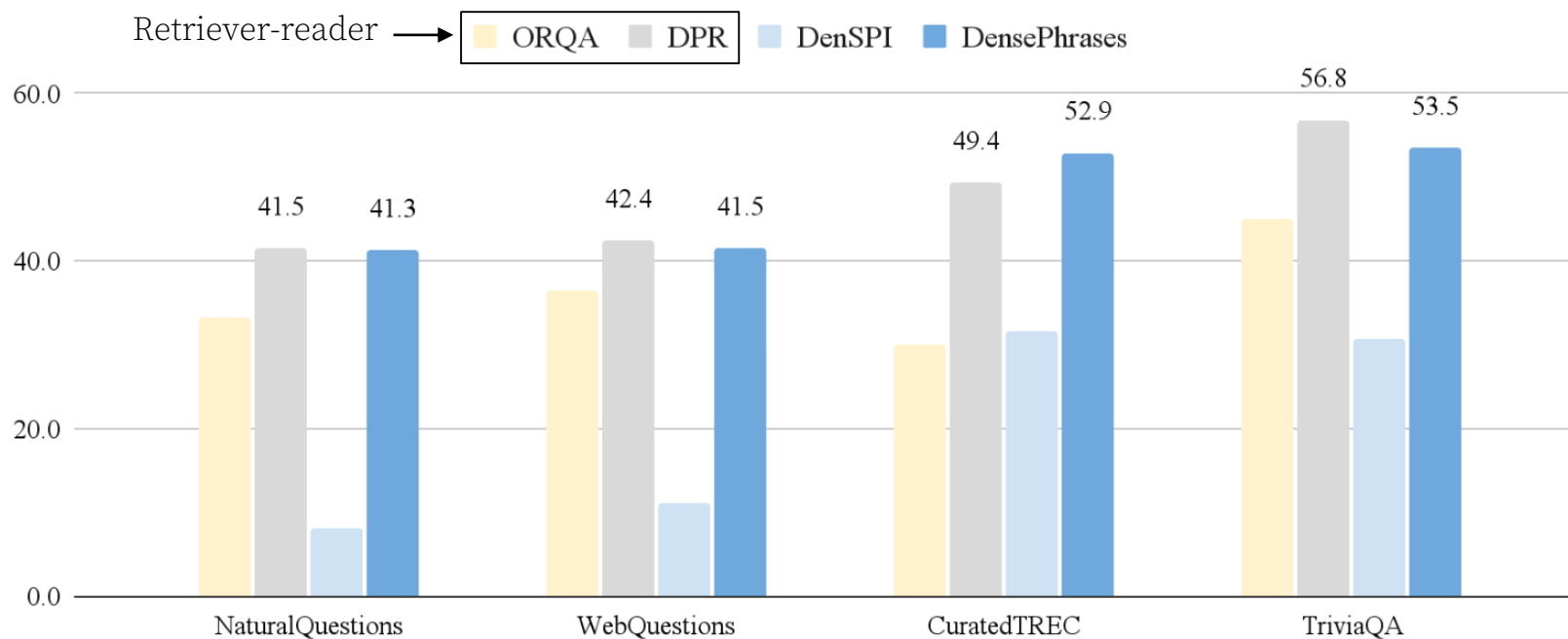
# 01

## Phrase Retrieval is **Accurate** and **Fast**



## 01

# Phrase Retrieval is **Accurate** and **Fast**

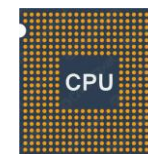


**Without any reader model**, phrase retrieval is **competitive** with retriever-reader approaches.

0.04 Q/sec  
(DPR)

<

13.6 Q/sec  
(DensePhrases)



**Dense phrase retrieval** makes open-domain QA **fast** and **simple**!





02

## Research Motivation

## 02

# Fixed Granularity for Text Retrieval

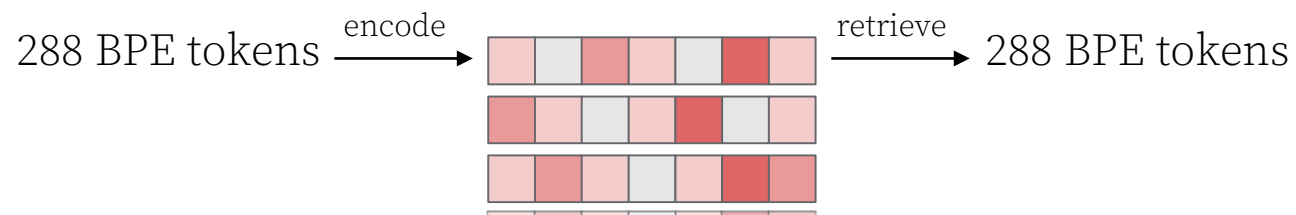
## Sentence Retrieval

SBERT (Reimers et al., 2019), SimCSE (Gao et al., 2021): **1 sentence**

## Passage Retrieval

ORQA (Lee et al., 2019): **288 BPE tokens** for a passage

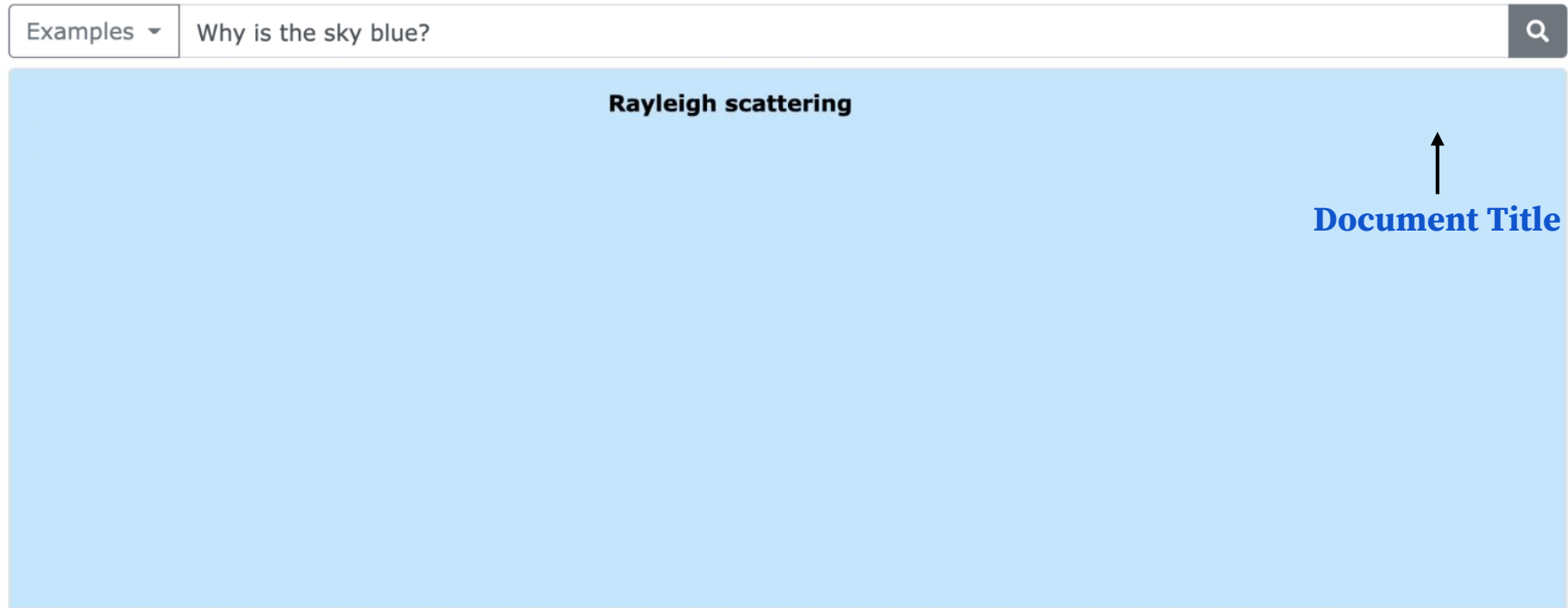
DPR (Karpukhin et al., 2020): **100 words** for a passage



**Different** index for **different** granularity?

## 02

# Phrases as a Basic Retrieval Unit

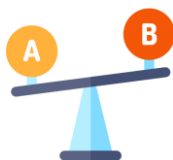


Retrieving **Phrases**  $\Rightarrow$  Sentences  $\Rightarrow$  Passages  $\Rightarrow$  Documents  $\Rightarrow$  ...

**Single** index for **multi** granularity!

# 02

## In This Talk ...



Q1: Is this **better** than passage retrievers?

Experiment #1: Passage Retrieval / Experiment #2: Open-domain QA



Q2: **Why** does this work?

Analysis / Experiment #3: Entity Linking & Dialogue



Q3: How **efficient** is this?

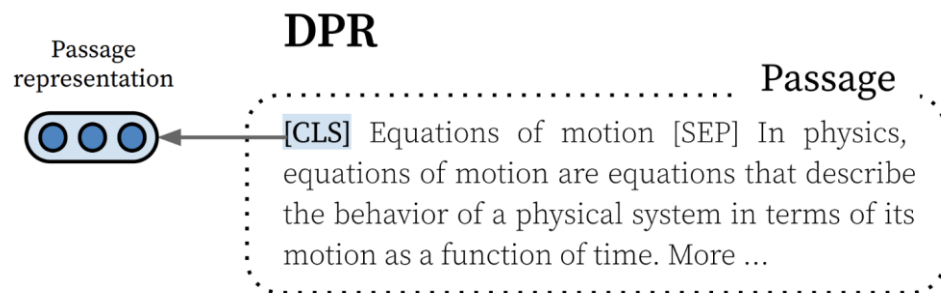
Phrase Filtering & Quantization-aware Fine-tuning



# 03

## Formulation / Experiments #1, #2

## Passage Retrieval

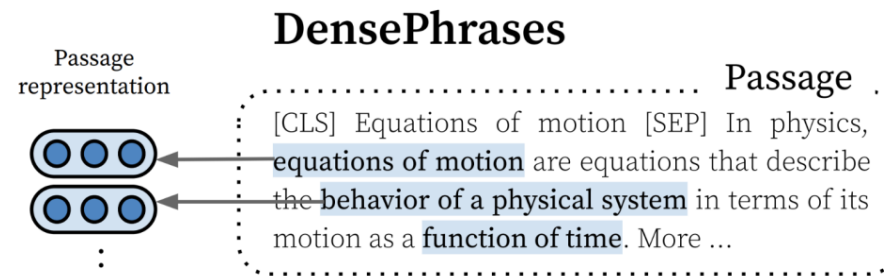


Question vector  
↓

$$f(p, q) = E_p(p)^\top E_q(q)$$

↑  
Single vector for each passage

## Phrase-based Passage Retrieval



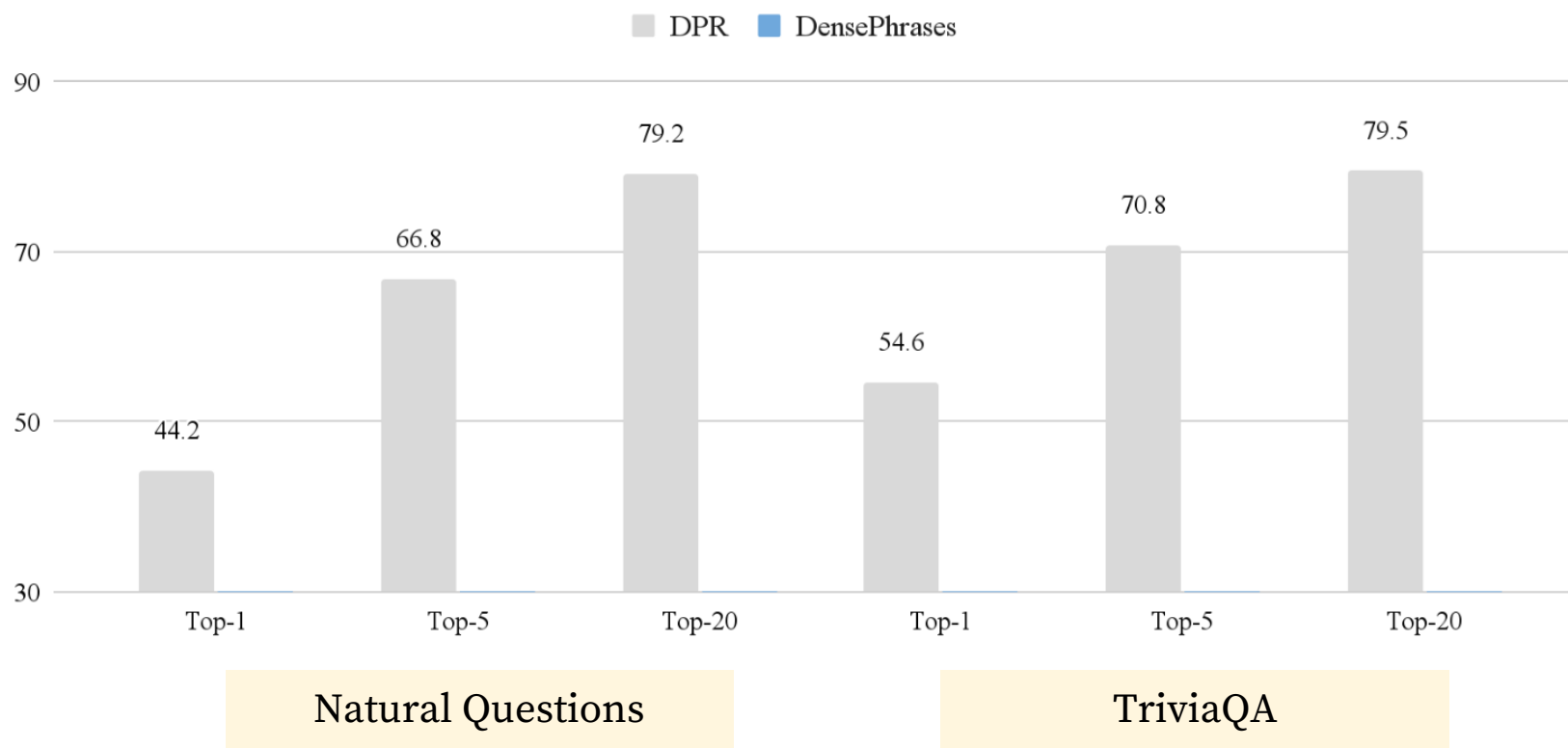
Phrase vector  
↓

$$\tilde{f}(p, q) := \max_{s^{(p)} \in \mathcal{S}(p)} E_s(s^{(p)})^\top E_q(q)$$

↑  
Multiple (phrase) vectors for each passage

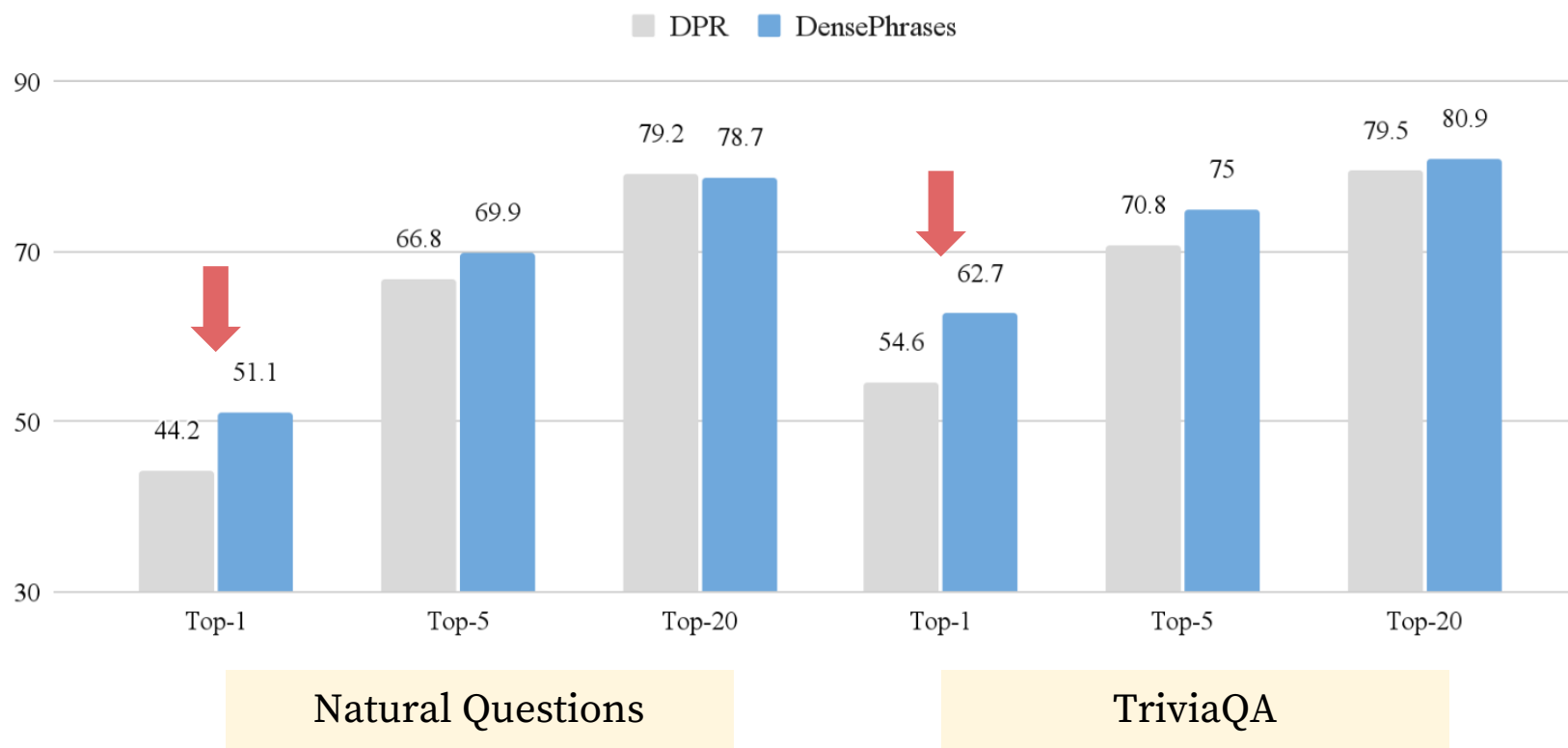
## 03

# Passage Retrieval: DPR vs DensePhrases



## 03

# Passage Retrieval: DPR vs DensePhrases



Without any re-training, **DensePhrases outperforms DPR** on passage retrieval!

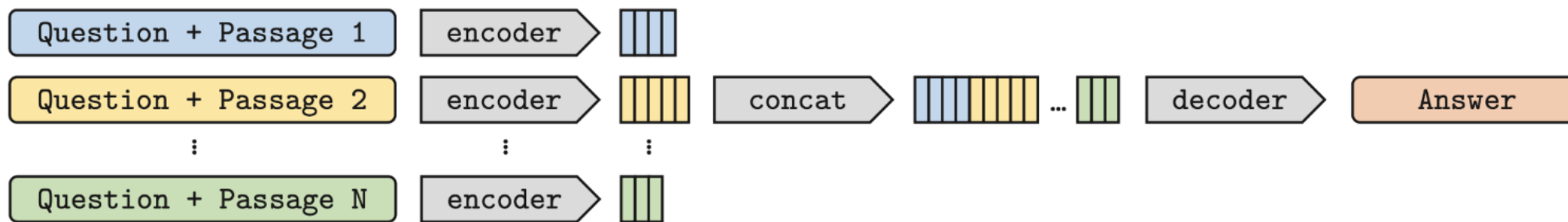
Larger gains when **k** is small.



## 03

# Fusion-in-Decoder for Open-domain QA

Izacard and Grave, 2021



Feeds top-k passages from **DPR** to **T5** (Raffel et al., 2020) to generate answers.

**FiD** achieves state-of-the-art performance **when k is large** (e.g., k=100).

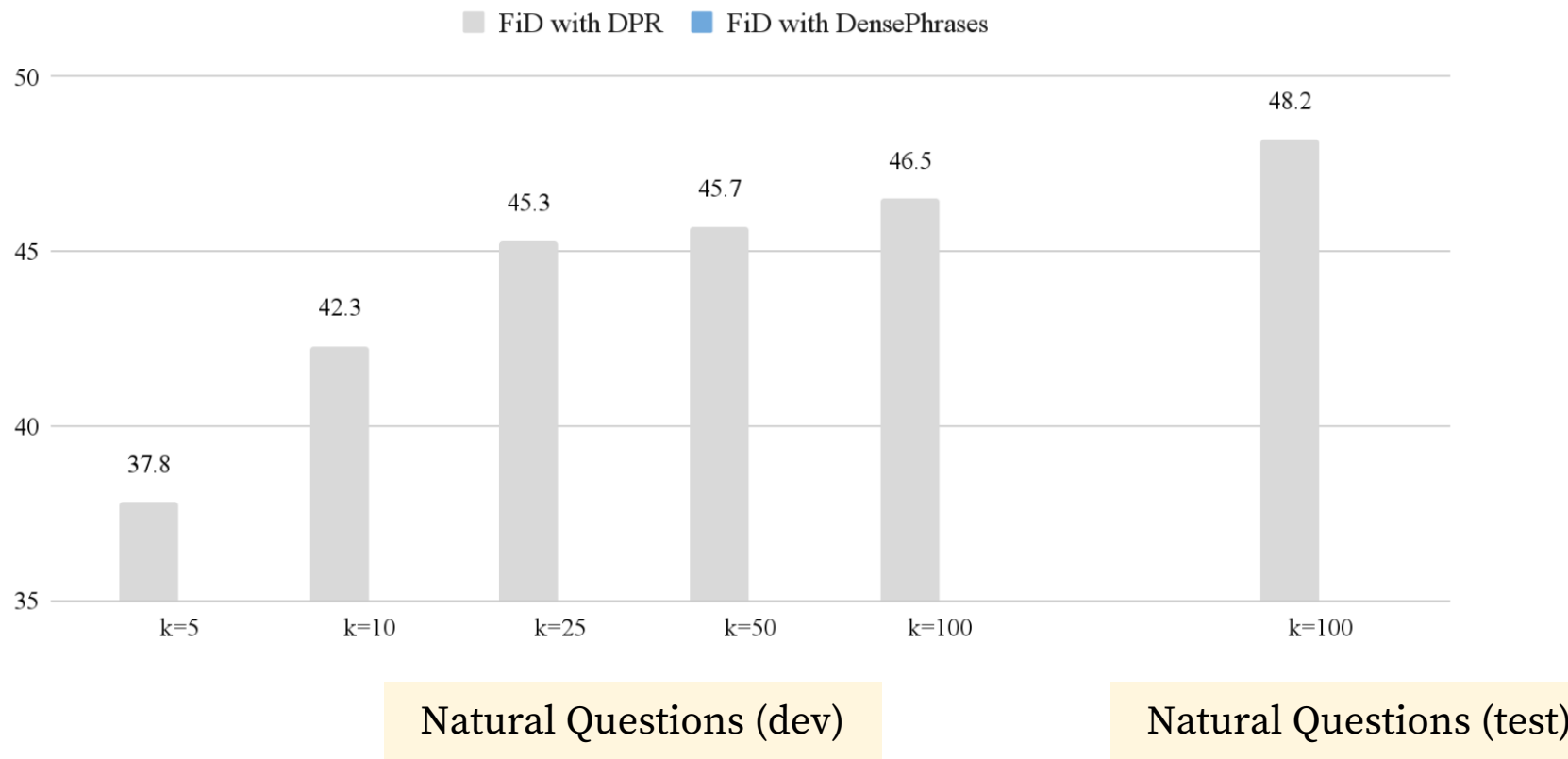
Requires 64 **32GB** V100 GPUs for training!



Feed top-k passages from **DensePhrases** to T5 to generate answers?

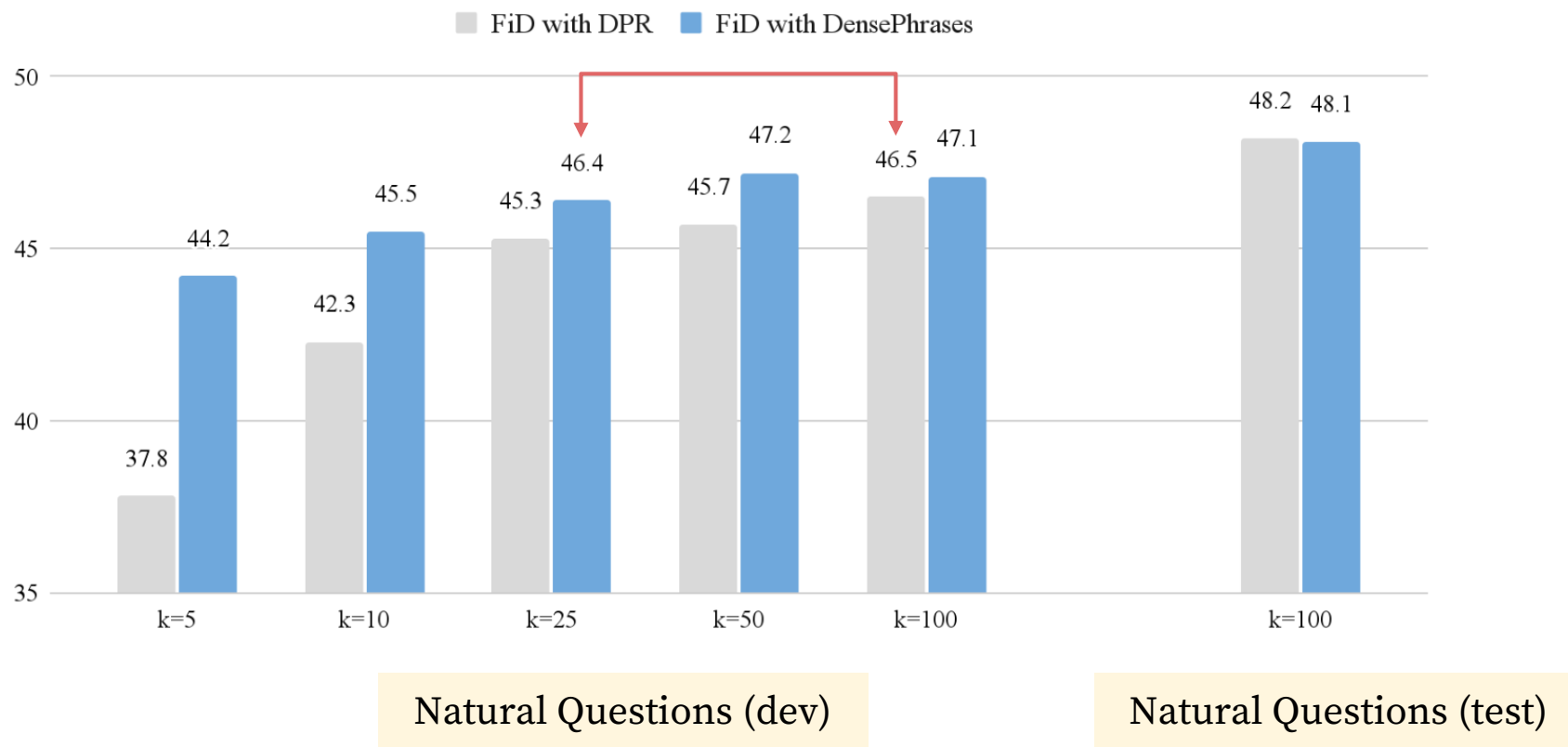
## 03

# Open-domain QA: DPR vs DensePhrases



## 03

# Open-domain QA: DPR vs DensePhrases



**DensePhrases outperforms DPR** on open-domain QA (+6.4 EM when k=5).

**k=25~50 is enough** for good performance ( $k \leq 50$  fits in **24GB**)

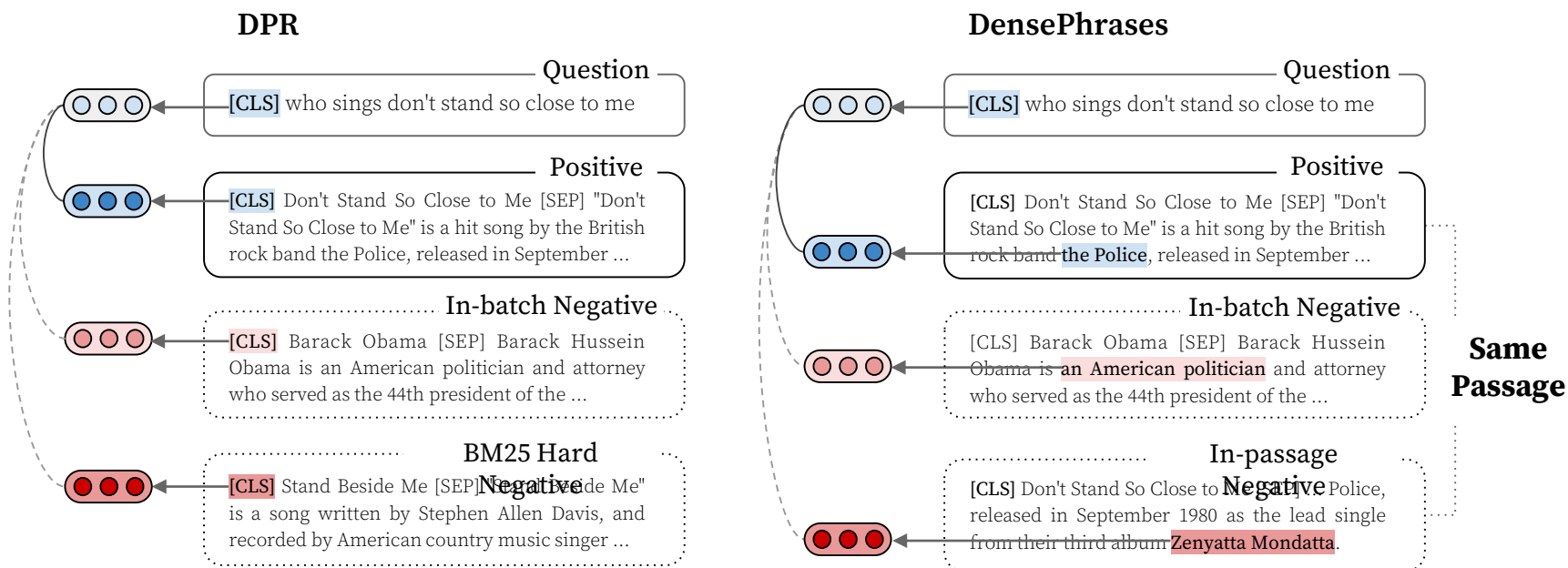


# 04

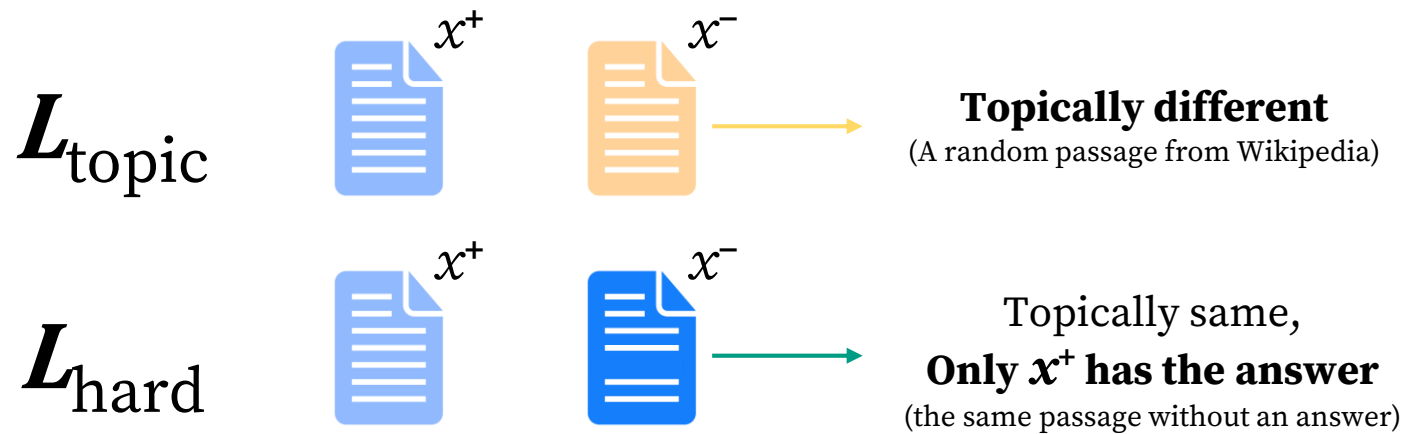
## **Analysis / Experiments #3**

## 04

# Why DensePhrases > DPR on Passage Retrieval?



**In-passage negatives** in DensePhrases work similar to **BM25 hard negatives** in DPR!



For both metrics, **lower numbers** are better.


**DPR** has good  $L_{\text{topic}}$  while **DensePhrases** has good  $L_{\text{hard}}$ .

## 04

 $L_{\text{topic}}$  and  $L_{\text{hard}}$  : What Do They Really Mean?

DPR (Karpukhin et al., 2020)

Where is Princeton University located? Run


Title: *Princeton University* → topically relevant! Retrieval ranking: #2  $P(p|q)=0.43$   $P(a|p,q)=0.94$   $P(q|p,q)=0.41$  

... "Cherokee Advocate", graduated in 1844. Princeton University Princeton University is a private Ivy League research university in **Princeton, New Jersey**. Founded in <http://qa.cs.washington.edu:2020/>

DensePhrases (Lee et al., 2021)

Examples ▾ Where is Princeton University located? Q

19 results (106ms) Real-time Search English Wikipedia (2018.12.20)

The New York metropolitan area is home to many prestigious institutions of higher education. Three Ivy League universities: Columbia University in Manhattan, New York City; Princeton University in **Princeton, New Jersey**; Yale University in New Haven, Connecticut – all ranked amongst the top 3 U.S. national  New York metropolitan area

<http://densephrases.korea.ac.kr>

Good  $L_{\text{hard}}$  can give correct answer even when **the passage is less relevant.**

topically less relevant, **but still correct answer!**



For many coarse-granularity retrieval,  
**we need good  $L_{\text{topic}}$  !**

### Entity Linking

*United Nations Security Council*

[START\_ENT] **Security Council** [END\_ENT] members expressed concern on Thursday.



### Knowledge-grounded Dialogue

*Yamaha Corporation*

Have you heard of Yamaha? They started as a piano manufacturer in 1887!

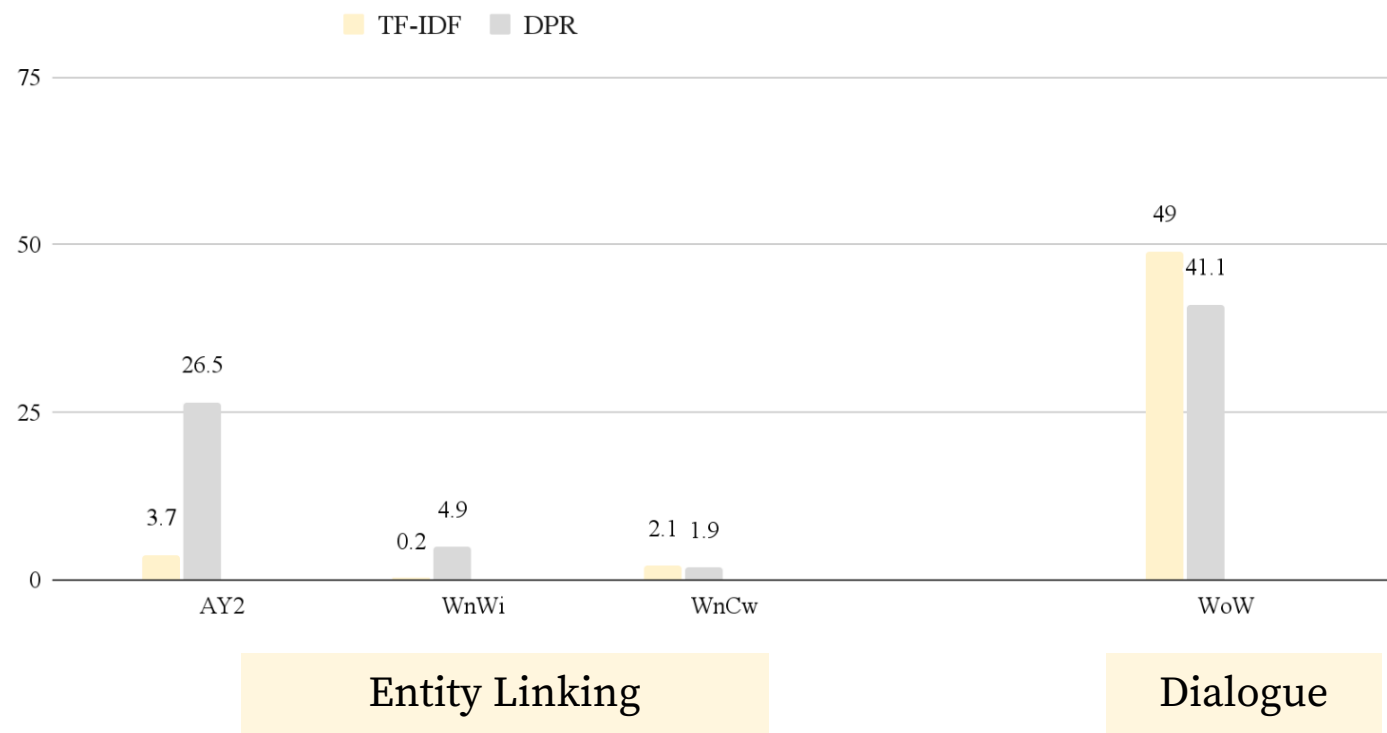


Only **one document** is relevant (annotated) for each query!  
(KILT; Petroni et al., 2021)




## 04

# Retrieval for Entity Linking & Dialogue



# Retrieval for Entity Linking & Dialogue

Examples ▾ do you like hip hop? 

20 results (93ms) ☒ Real-time Search English Wikipedia (2018.12.20)

75 — Radio DJs or radio personalities introduce and play music that is broadcast on AM, FM, digital or Internet radio stations. Club DJs, commonly referred as DJs in general, play music at musical events, such as parties at music venues or bars, music festivals, corporate and private events. Typically, club DJs mix music recordings from two or more sources using different mixing techniques in order to produce non-stopping flow of music. One key technique used for seamlessly transitioning from one song to another is beatmatching. A DJ who mostly plays and mixes one specific music genre is often given the title of that genre; for example, a DJ who plays **hip hop music** is called a hip hop DJ, a DJ who plays house music is a house DJ, a DJ who plays techno is called a techno DJ, and so on. The quality of a DJ performance (often called a DJ mix or DJ set) consists of two main features: technical skills, or how well can DJ operate the equipment and produce smooth transitions between two or more recordings and a playlist, or ability of a DJ to select most suitable recordings also known as "reading the crowd". Disc jockey

50 —  $f(s|D,q)=90.02$

25 — Hip-hop music has reached the cultural corridors of the globe and has been absorbed and reinvented around the world. Hip hop music expanded beyond the US, often blending local styles with hip hop. Hip hop has globalized into many cultures worldwide, as evident through the emergence of numerous regional scenes. It has emerged globally as a movement based upon the main tenets of hip hop culture. The music and the art continue to embrace, even celebrate, its transnational dimensions while staying true to the local cultures to which it is rooted. Hip-hop's impact differs depending on each culture. Still, the one thing virtually all hip hop artists worldwide have in common is that they acknowledge their debt to those **African-American people in New York** who launched the global movement. Hip hop music → **topically relevant!**  
(annotated)

0 — Hip-hop music has reached the cultural corridors of the globe and has been absorbed and reinvented around the world. Hip hop music expanded beyond the US, often blending local styles with hip hop. Hip hop has globalized into many cultures worldwide, as evident through the emergence of numerous regional scenes. It has emerged globally as a movement based upon the main tenets of hip hop culture. The music and the art continue to embrace, even celebrate, its transnational dimensions while staying true to the local cultures to which it is rooted. Hip-hop's impact differs depending on each culture. Still, the one thing virtually all hip hop artists worldwide have in common is that they acknowledge their debt to those African-American people in **New York who launched the global movement.** Hip hop music → **topically relevant!**  
(annotated)

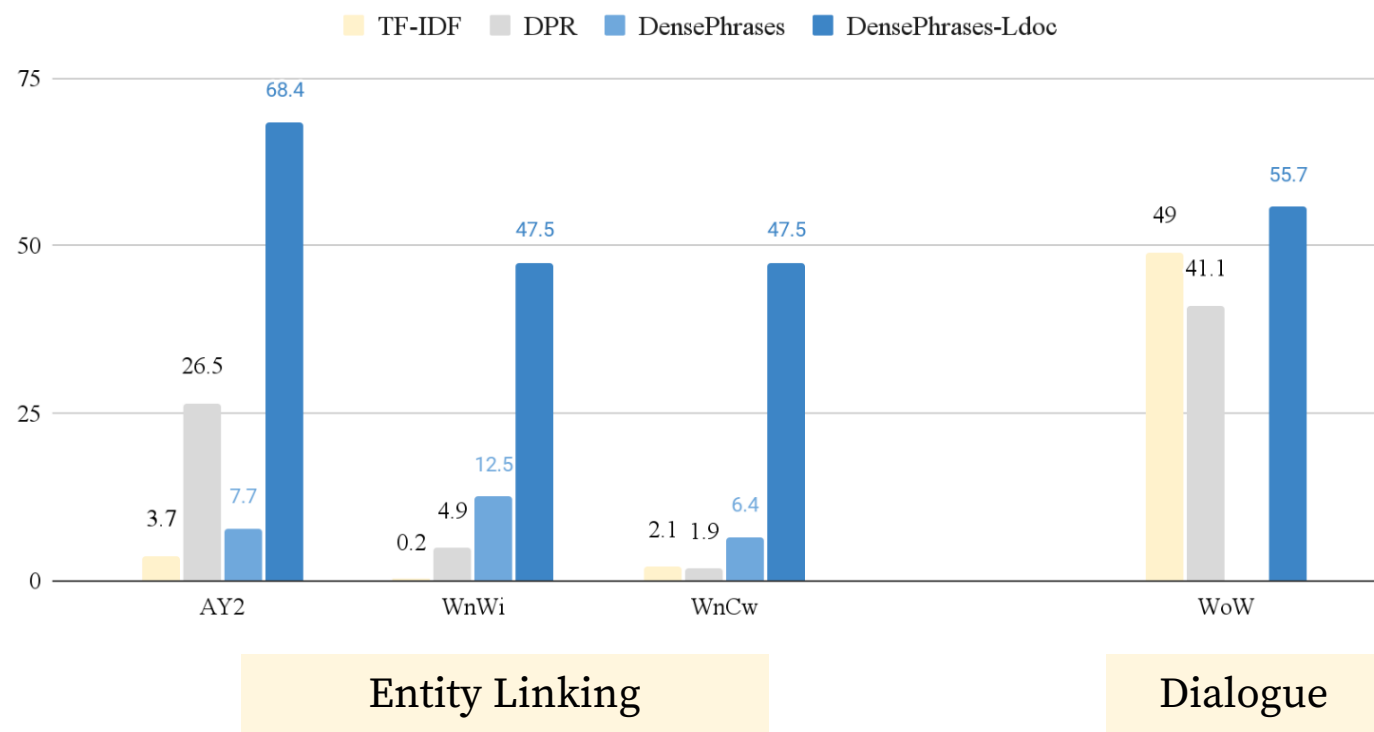
$f(s|D,q)=88.63$

$f(s|D,q)=88.23$

Maximize the marginal probability of  
**any phrases in the relevant document**

## 04

# Retrieval for Entity Linking & Dialogue



DensePhrases can be adapted to **retrieve topically relevant documents!**

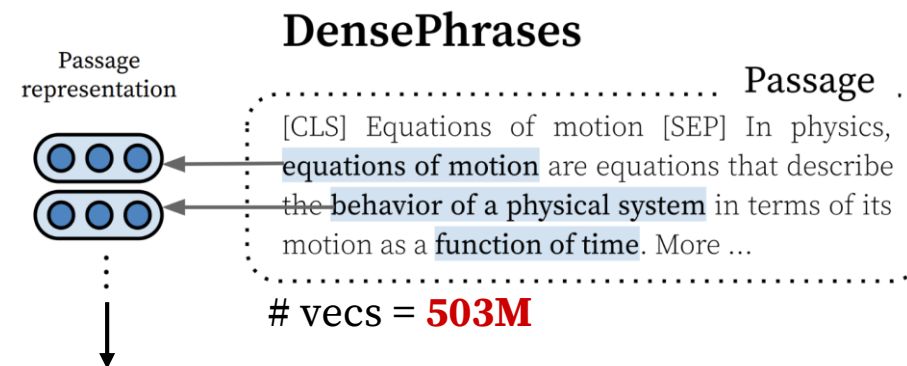
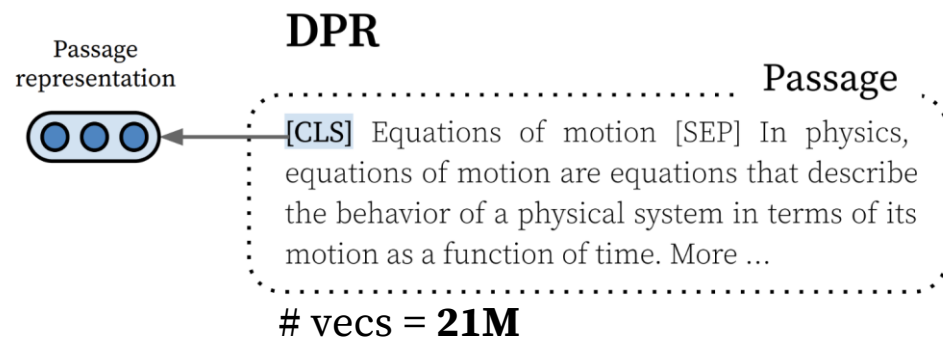


05

# Complexity Analysis

# Problem of Multi-vector Encoding

Luan et al., 2021; Khattab and Zaharia, 2020



More vectors, **more space!**



Phrase indexes are **heavy!**



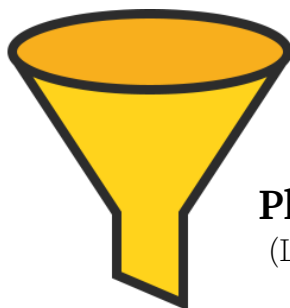
**1.2TB** (Seo et al., 2019)

**1.5TB** (Lee et al., 2020)

**320GB** (Lee et al., 2021)

# Reducing the Size of Phrase Index

“The New York metropolitan area is home to many prestigious institutions of higher education.”



**Phrase Filter**

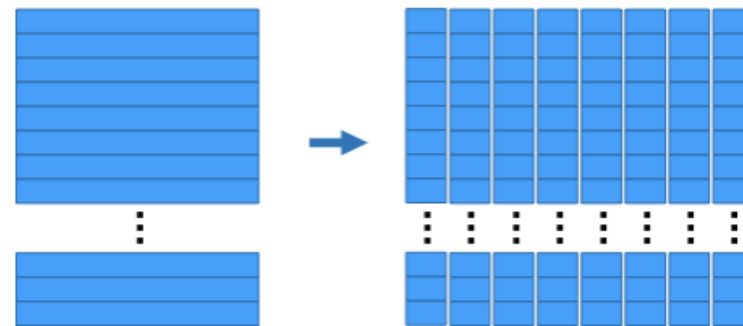
(Lee et al., 2021)

“**The New York metropolitan area**”

“prestigious institutions”

“higher education”

...



**Optimized Product Quantization**

(Ge et al., 2013)

+

**Query-side Fine-tuning**

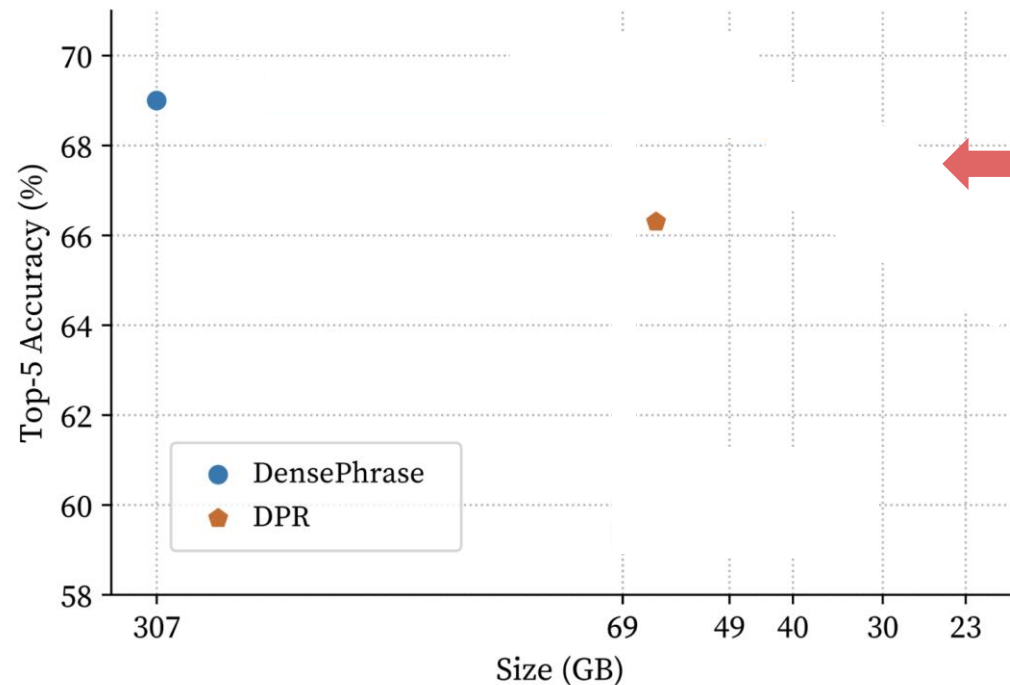
(Lee et al., 2021)

=

**Quantization-aware Fine-tuning**

## 05

# Reducing the Size of Phrase Index



We can safely reduce the size down to **23GB!** (DPR = 69GB)

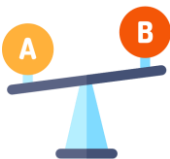
DensePhrases with **# vector/passage = 8.8** is similar to DPR.



06

**Conclusion**





Q1: Is this **better** than passage retrievers?

Yes! **DensePhrases** > **DPR** on passage retrieval and open-domain QA!



Q2: **Why** does this work?

Better at **fine-grained entailment**, can be used for coarse retrieval.



Q3: How **efficient** is this?

Can safely reduce the index size from **307GB to 23GB!**

Paper: <https://arxiv.org/abs/2109.08133>

Code & Models: <https://github.com/princeton-nlp/DensePhrases>

Demo: <http://densephrases.korea.ac.kr/>

E-mail: [jinkyuklee@cs.princeton.edu](mailto:jinkyuklee@cs.princeton.edu)

# 감사합니다

Phrase Retrieval Learns Passage Retrieval, Too

