

End-to-End Training of Neural Retrievers for Open-Domain Question Answering

ACL 2021

- Devendra Singh Sachan
- PhD student at Mila and McGill University
- Work done during internship at NVIDIA



Devendra Sachan



Mostofa Patwary



Mohammad Shoeybi



Neel Kant



Wei Ping



William Hamilton



Bryan Catanzaro

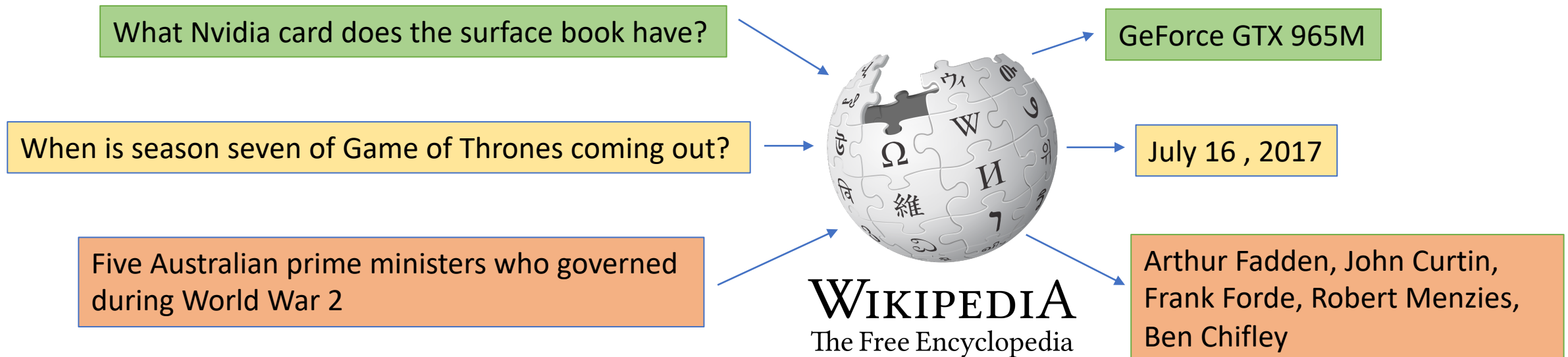
Background and Problem Statement

Retriever Pre-training

End-to-End Supervised Training

Problem Setup: Open-Domain QA

- **Input:** Question and evidence documents such as Wikipedia (millions of documents)
- **Output:** Answer



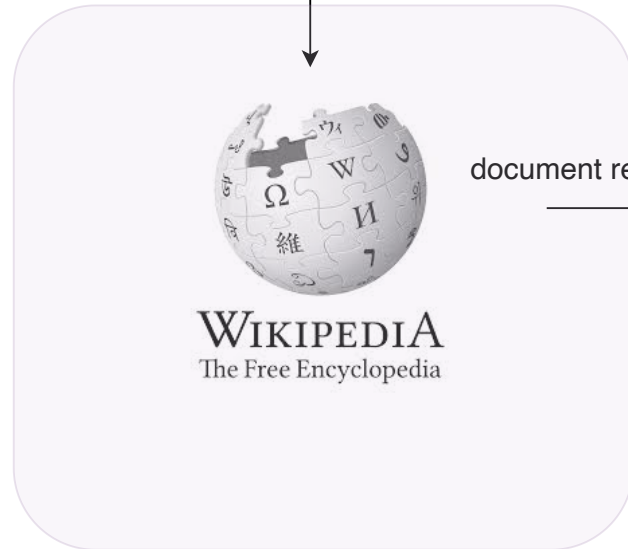
Background: Open-Domain QA

- Two-stage approach

1. Document retrieval from evidence

2. Answer Extraction

What Nvidia card does the surface book have?



Information Retrieval

Ex: TFIDF, BM-25

Surface Book
From Wikipedia, the free encyclopedia

The **Surface Book** is a 2-in-1 PC designed and produced by Microsoft, part of the company's Surface line of personal computing devices. Surface Book is distinguished from other Surface devices primarily by its full-sized, detachable keyboard, which uses a dynamic fulcrum hinge that expands when it is opened. The keyboard contains a second battery, a number of ports and an optional discrete graphics card used when the screen part, also dubbed as the clipboard by Microsoft, is docked to it. Contrary to Surface Pro devices, which are marketed as tablets, the Surface Book is marketed as a laptop. Microsoft's first device marketed as such.

Contents [hide]

- 1 History
- 2 Features

Nvidia
From Wikipedia, the free encyclopedia

For the screen reader known as "NVDA", see *NonVisual Desktop Access*.

Nvidia Corporation^{mw} [ⓘ] (^{en}/^{vidi}/ ^{en-vid-ee}-ə) is an American multinational technology company incorporated in Delaware and based in Santa Clara, California.^[P] It designs graphics processing units (GPUs) for the gaming and professional markets, as well as system on a chip units (SoCs) for the mobile computing and automotive market. Its primary GPU product line, labeled "GeForce", is in direct competition with Advanced Micro Devices' (AMD) "Radeon" products. Nvidia expanded its presence in the gaming industry with its handheld *Shield Portable*, *Shield Tablet*, and *Shield Android TV* and its cloud gaming service GeForce Now.

In addition to GPU manufacturing, Nvidia provides parallel processing capabilities to researchers and scientists that allow them to efficiently run high-performance applications. They are deployed in supercomputing sites around the world.^[R] More recently, it has moved into the mobile computing market, where it produces Tegra mobile processors for smartphones and tablets as well as vehicle navigation and entertainment systems.^{[R][P]} In addition to AMD, its competitors include Intel and Qualcomm.

Contents [hide]

- 1 History
 - 1.1 Major releases and acquisitions
 - 1.2 Class action lawsuit
 - 1.3 Apple/Nvidia web driver controversy
 - 1.4 Hardware Unboxed controversy
- 2 Finances
- 3 GPU Technology Conference
- 4 Product families
- 5 Open-source software support
- 6 Deep learning
 - 6.1 DGX
- 7 Inception Program
 - 7.1 2018 winners^[17]
 - 7.2 2017 winners^[17]
- 8 See also
- 9 Notes
- 10 References
- 11 External links

History [edit]

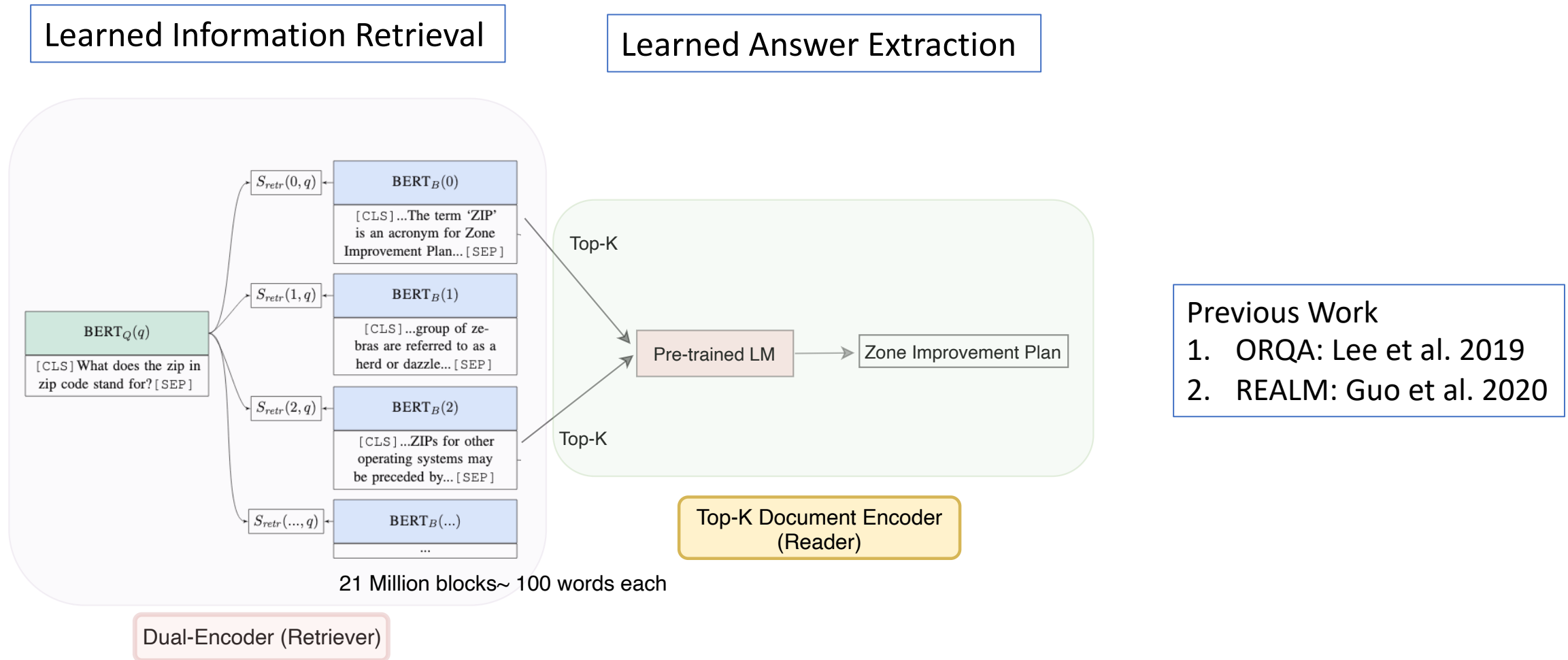
Nvidia was founded on April 5, 1993,^{[R][P]} by Jensen Huang (CEO as of 2020), a Taiwanese American, previously director of CoreWare at LSI Logic and a microprocessor designer at Advanced Micro Devices (AMD), Chris Malachowky, an electrical engineer who worked at Sun Microsystems, and Curtis Priem, previously a senior staff engineer and graphics chip designer at Sun Microsystems.

In 1993, the three co-founders believed that the proper direction for the next wave of computing was accelerated or graphics-based computing because it could solve problems that general-purpose computing could not. They also observed that video games were simultaneously one of the most computationally challenging problems and would have incredibly high sales volume. The two conditions don't happen very often. Video games became the company's flywheel to reach large markets and funding huge R&D to solve massive computational problems. With only \$40,000 in the bank, the company was born.^[11] The company subsequently received \$20 million of venture capital funding from Sequoia Capital and others.^[12] Nvidia initially had no name and the co-founders named all their files NV, as in "next version". The need to incorporate the company prompted the co-founders to review all words with those two letters, leading them to "nvidia", the Latin word for "envy".^[11] Nvidia went public on January 22, 1999.^{[13][14]}

Answer Extraction

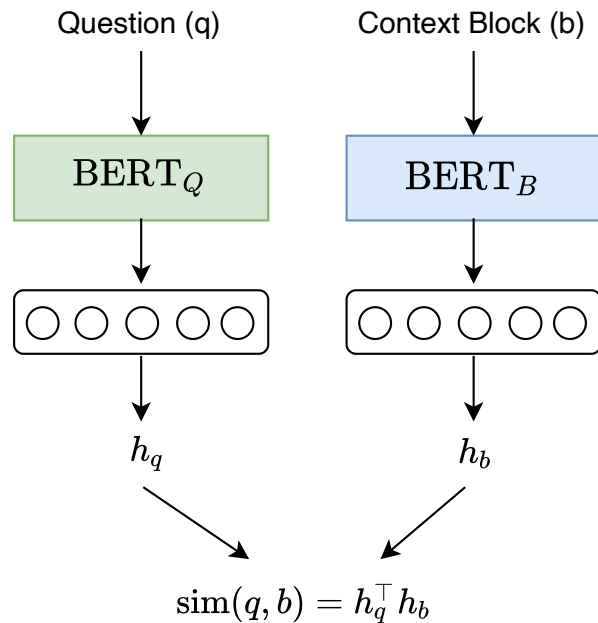
GeForce GTX 965 M

Background: Neural Models for Open-Domain QA



Prior Work: Learned Information Retrieval

- **Retriever:** Dual-encoder model
- Train from query-context pairs



$$D = q_i, \text{ Query}$$
$$b_i^+, \text{ Positive Context}$$
$$b_j^-, \text{ Other Context}$$

$$\mathcal{L} = -\log \frac{e^{\text{sim}(q_i, b_i^+)}}{e^{\text{sim}(q_i, b_i^+)} + \sum_{j=1}^n e^{\text{sim}(q_i, b_j^-)}}$$

Supervised Training: Dense Passage Retriever (DPR)

- **Positive Examples:**

- Included with the question-answering datasets.
- Top-ranked BM25 passages in Wikipedia containing the answer string.

- **Negative Examples:**

- **Hard negatives:** Passages of high BM25 scores that **DO NOT** contain the answer.
- **In-batch negatives:** Positive passages of **OTHER** questions.

Background and Problem Statement

Retriever Pre-training

End-to-End Supervised Training

Proposed Approach

- Scale the retrieval similarity score by *square root of hidden size*.

$$\text{sim}(q, b) = \frac{h_q^\top h_b}{\sqrt{d}}$$

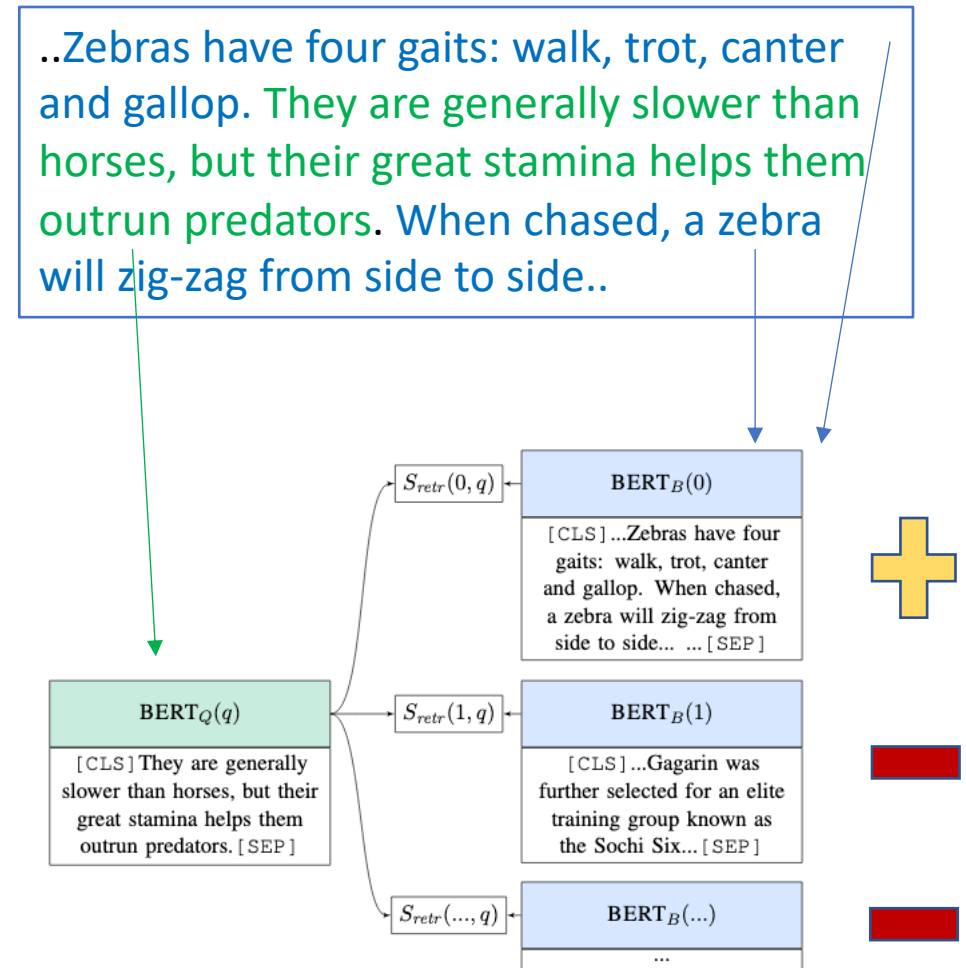
- As model dimensions increases, similarity score also increases
 - *Hypothesis*: scaled score leads to a better optimization
- Perform *longer supervised training* of the retriever.

Proposed Approach

- Unsupervised pre-training + Supervised training of retriever
 1. *Inverse Cloze Task* + DPR
 2. *Masked Salient Spans* + DPR

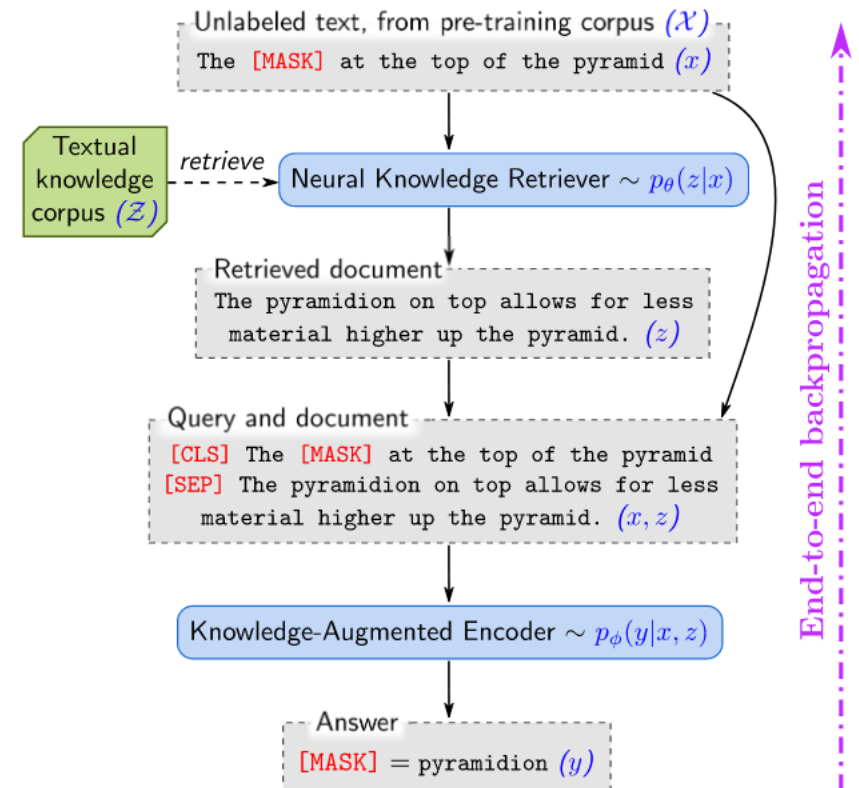
Retriever Pre-training by Inverse Cloze Task

- **Inverse Cloze Task (ICT)** – first proposed in *Lee et al., 2019*.
- Sample a sentence from a paragraph.
- Sentence can be considered as the *query*.
- Remaining sentences can be considered as the *context*.
- **Unsupervised** - can use all Wikipedia to train the model.



Retriever Pre-training by Masked Salient Spans

- **Masked Salient Spans (MSS)** training: first proposed in Guu et al, 2020 .
- Model: Retriever + Top-K Encoder
 - **Retriever**: Initialize with **ICT**.
 - **Top-K Encoder**: Initialize with **T5**.
- **Query**: masked named entities in a sentence.
- **Task**: generate masked spans conditioned on query + retrieved doc.



Experimental Setup

Datasets

- Natural Questions (NQ)
 - Collection of real questions by Google
 - Questions have short answers (< 5 words)
- TriviaQA
 - Collection of trivia questions

Dataset	Train	Val	Test
NQ	79,168	8,757	3,610
TriviaQA	78,785	8,837	11,313

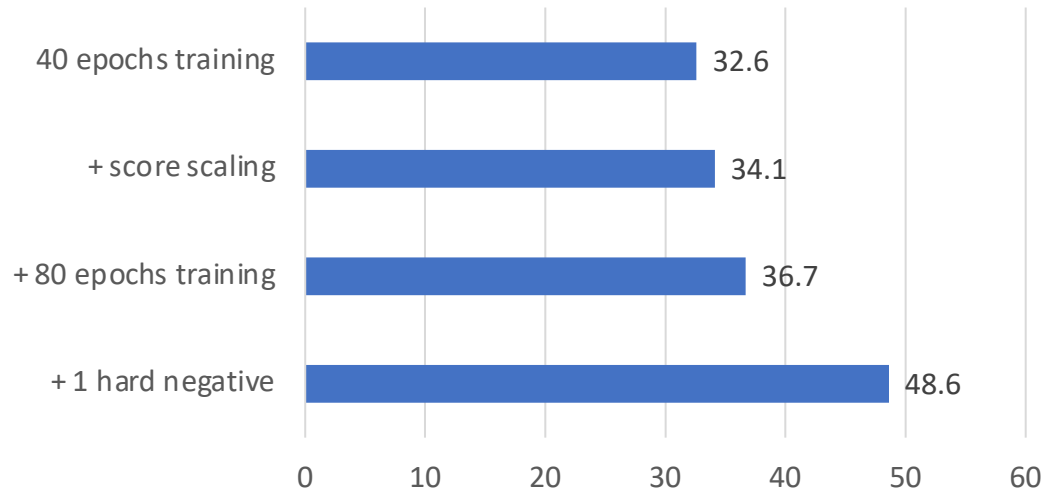
Evaluation Metric

Precision@top-K

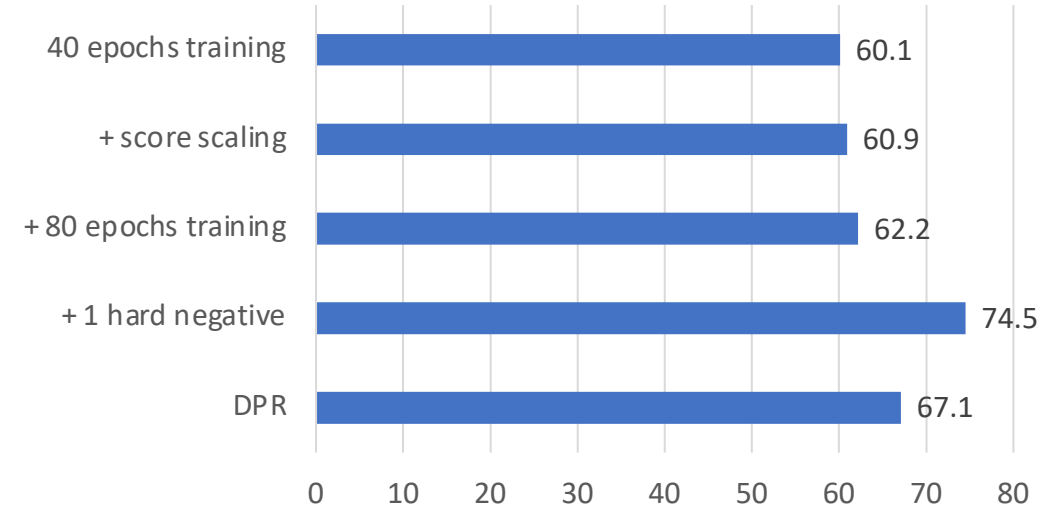
- Exact Match if answers exists in top-K documents or not

Results: Effect of Score Scaling and Longer Training

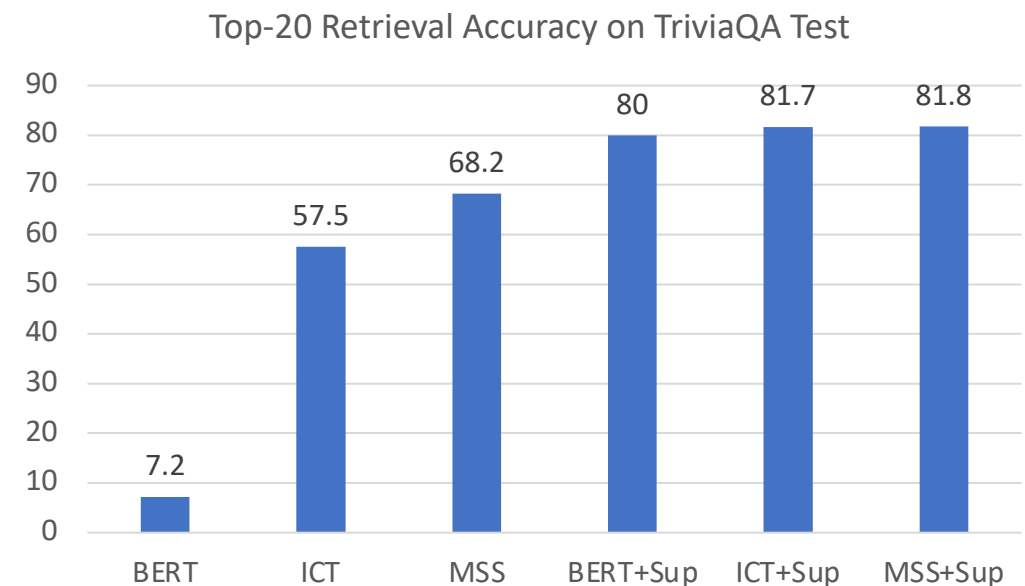
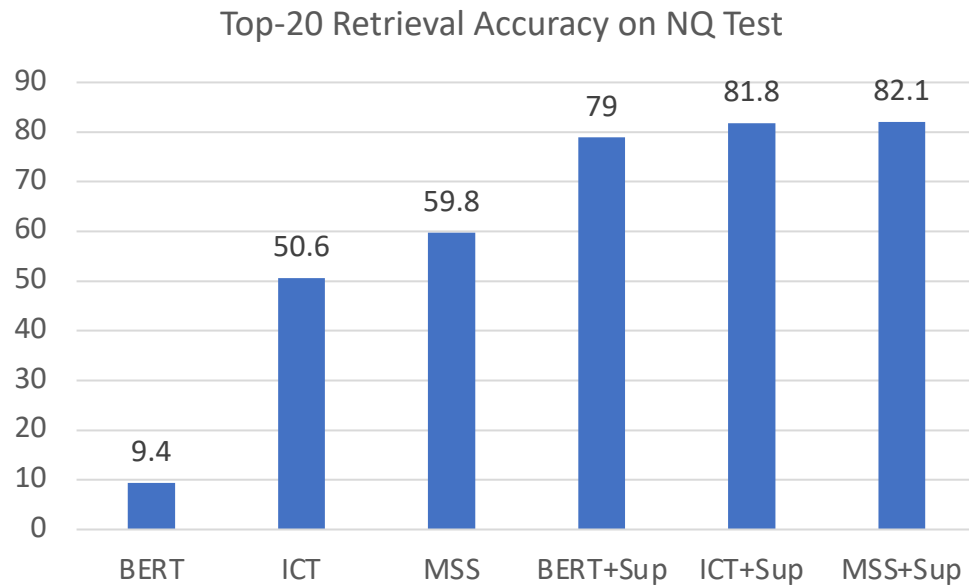
Top-1 Accuracy on NQ Test



Top-5 Accuracy on NQ Test



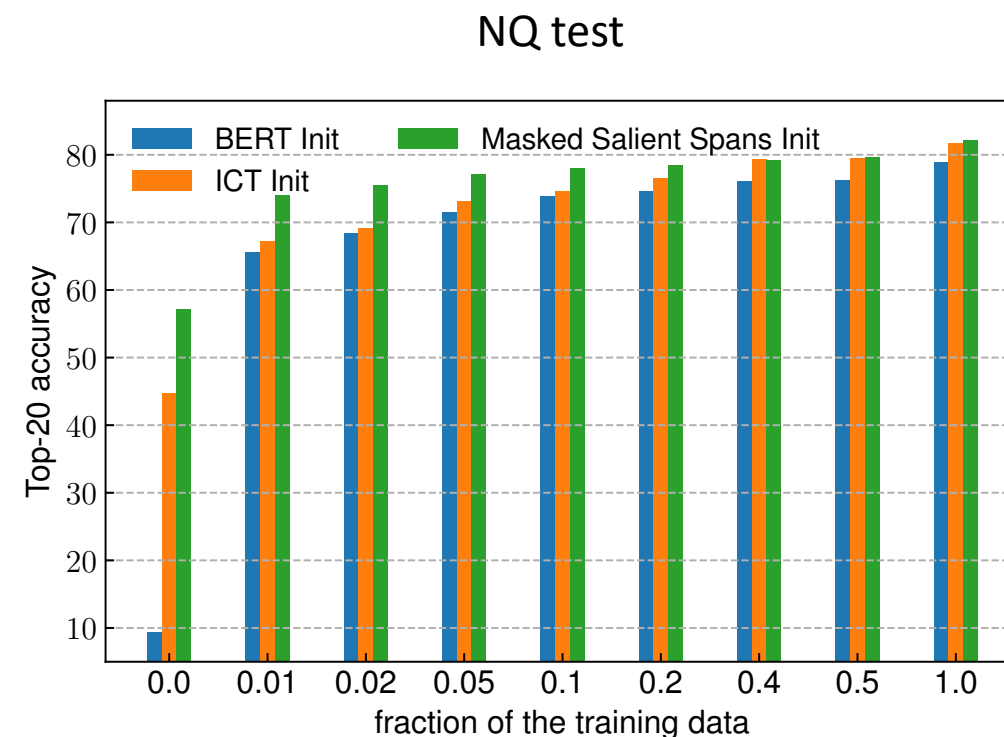
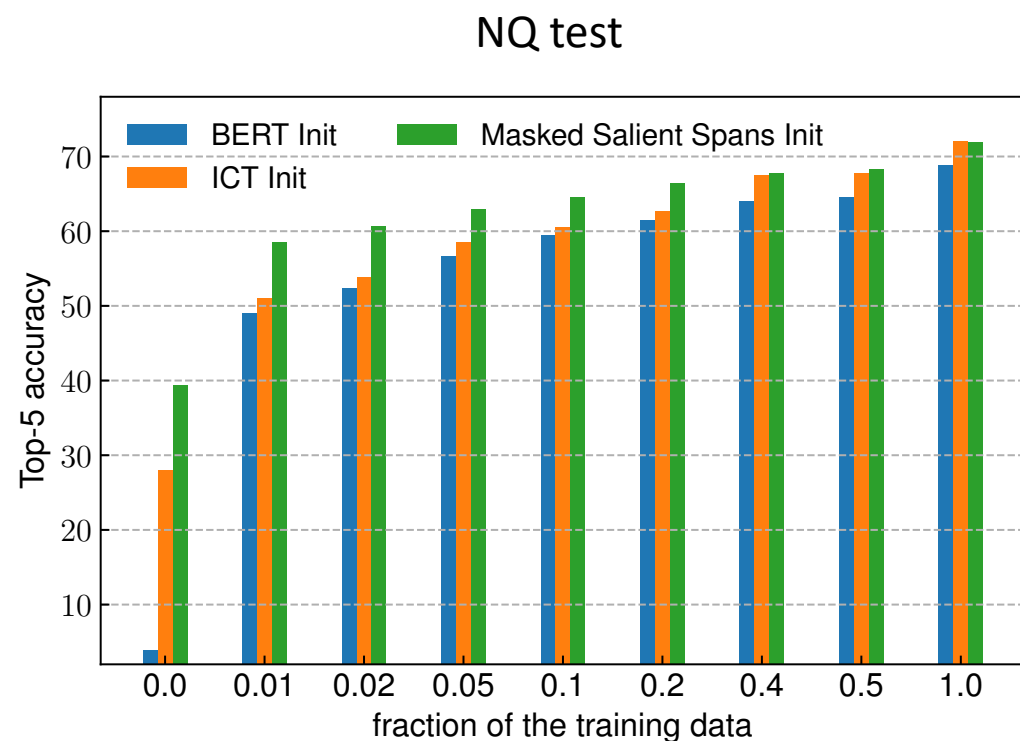
Results: Effect of Unsupervised Pre-training



ICT + Supervised and MSS + Supervised outperform Supervised retriever training

New state-of-the-art results!

Retrieval Accuracy: Effect of Amount of Training Data



1. MSS pre-training is more effective than ICT for lower-resource training data.
2. For high-resource setup, gains from MSS pre-training saturates to that of ICT pre-training.

Background and Problem Statement

Retriever Pre-training

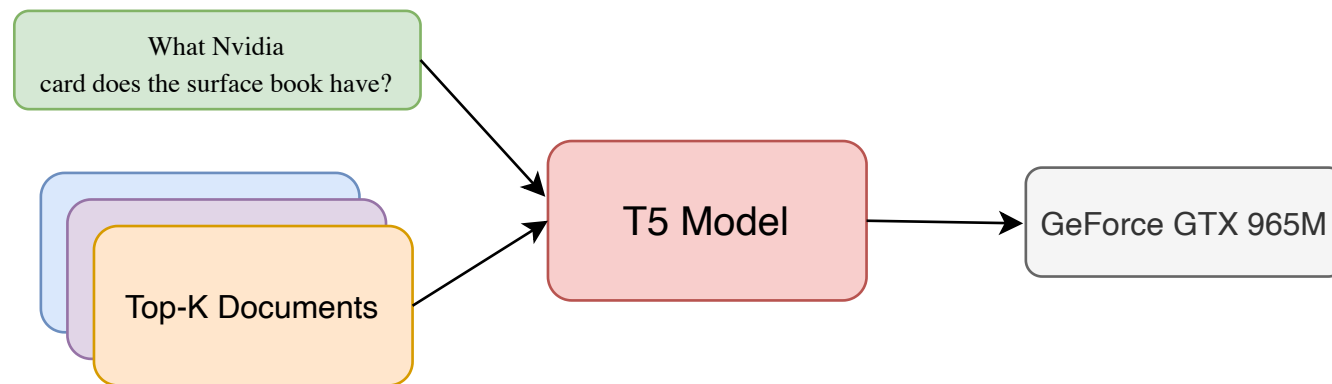
End-to-End Supervised Training

Neural Reader for Answer Extraction

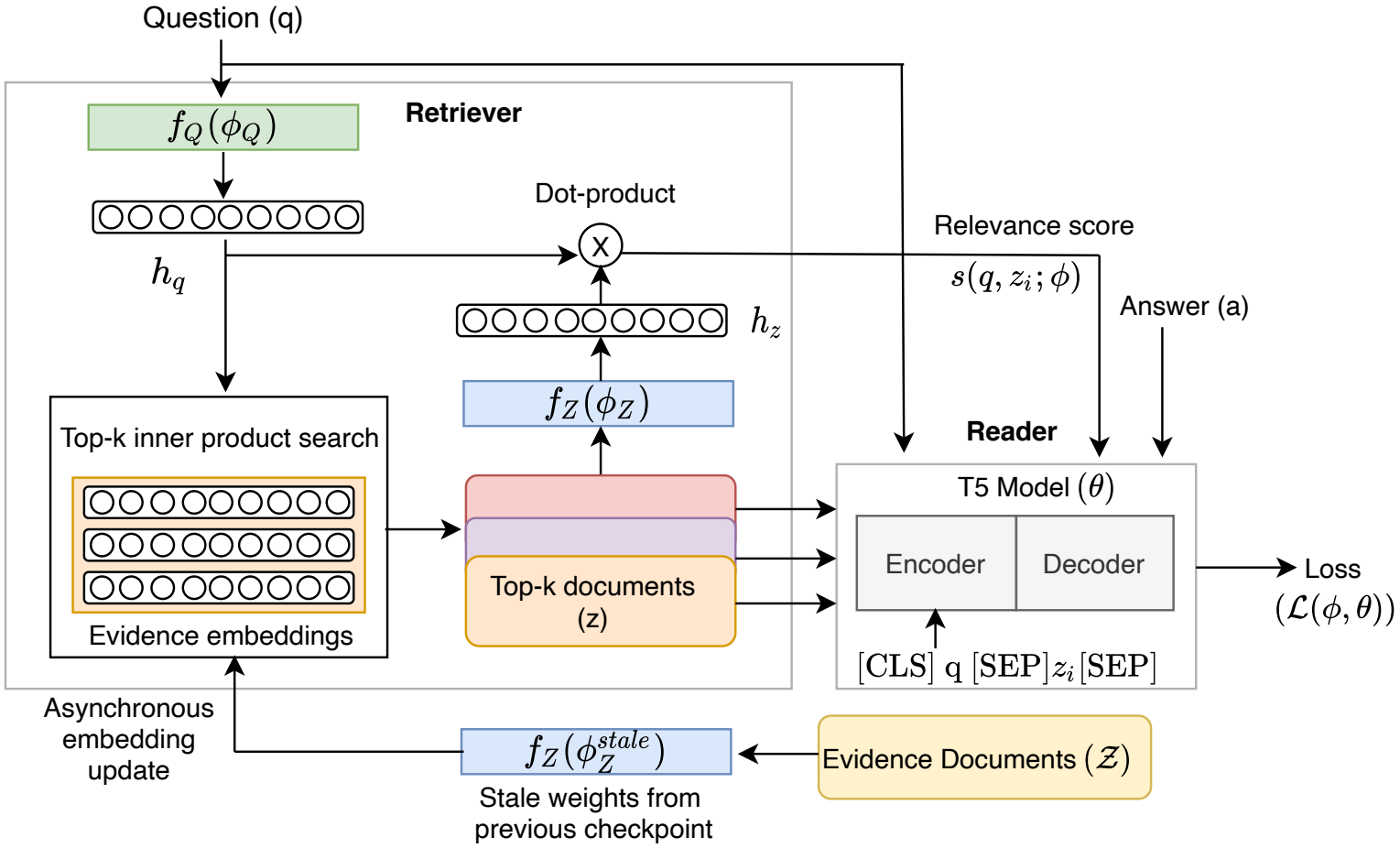
1. Retrieve top-K documents from **retriever**

$$\mathcal{K} = \underset{z_i \in \mathcal{Z}}{\operatorname{arg\,sort}} s(q, z_i; \phi)[:, k]$$

2. Encode top-K documents with **T5 seq-to-seq** model



End-to-End Supervised Training using QA Pairs



Approaches to Encode Top-K Documents

1. **Individual Top-K:** Encode each top-K document separately
2. **Joint Top-K:** Jointly encode all top-K documents

Retriever: ICT + DPR
Reader: pre-trained T5

Approach 1: Individual Top-K

Objective function: similarity weighted likelihood of each top-K document

$q = \text{question}, a = \text{answer}, z = \text{top-K doc}$

$$p(a | q) = \sum_k p(a | q, z_k) p(z_k | q)$$

Computed from T5

Computed from Retriever

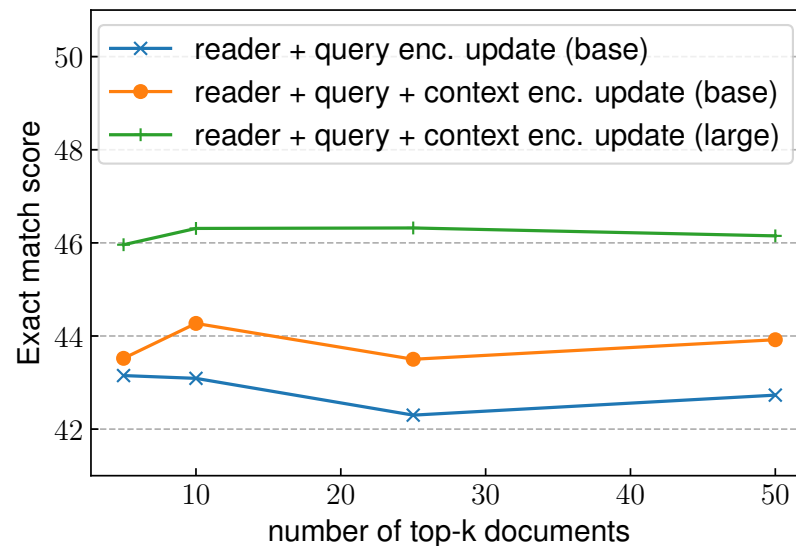
Decoding: Select the best answer among K answers

Results: Individual Top-K

Model	NQ	TriviaQA
<i>Base Configuration</i>		
ORQA	33.3	45.0
REALM	40.4	—
DPR	41.5	56.8
Individual Top-k	45.9	56.3
<i>Large Configuration</i>		
RAG	44.5	56.8
Individual Top-k	48.1	59.6

Better Retriever + T5 top-K encoder helps!

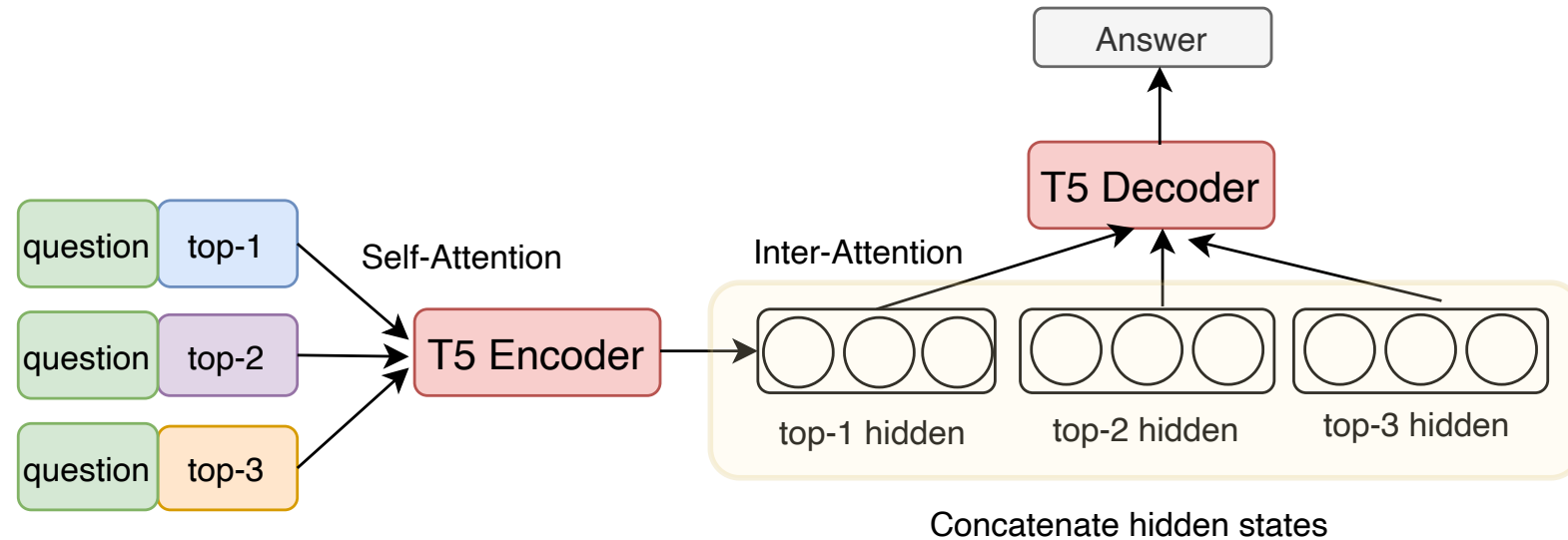
Effect of Async Retriever Update on NQ Dev



Context embedder update helps!

Approach 2: Joint Top-K

- **T5 Encoder:** compute hidden states of each top-K separately
- **T5 Decoder:** jointly attend to the concatenated hidden states



$$\text{attention}(q, a) \propto Q(a)K(x_1 \dots x_k; q) + \beta p(x_i | q)$$

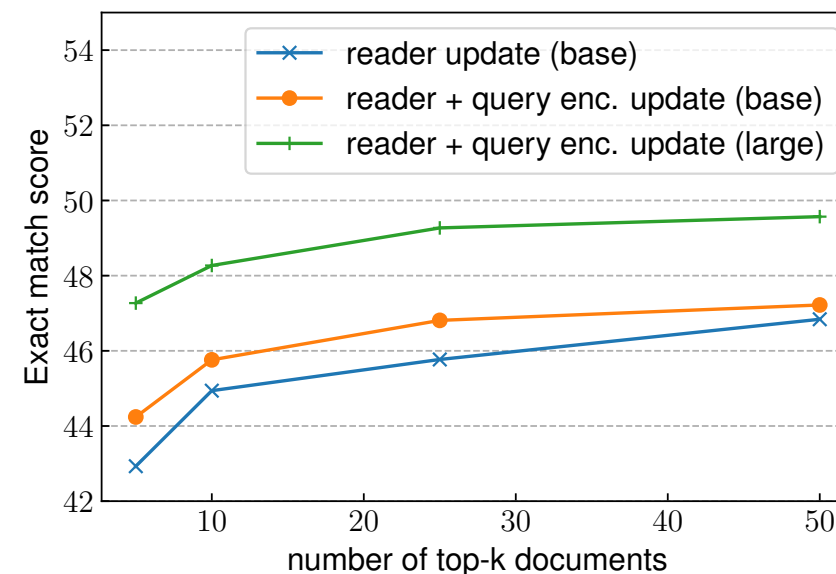
← Retriever similarity score bias

Results: Joint Top-K

Model	NQ	TriviaQA
<i>Base Configuration</i>		
FiD	48.2	65.0
Joint Top-k	49.2	64.8
<i>Large Configuration</i>		
FiD	51.4	67.6
Joint Top-k	51.4	68.3

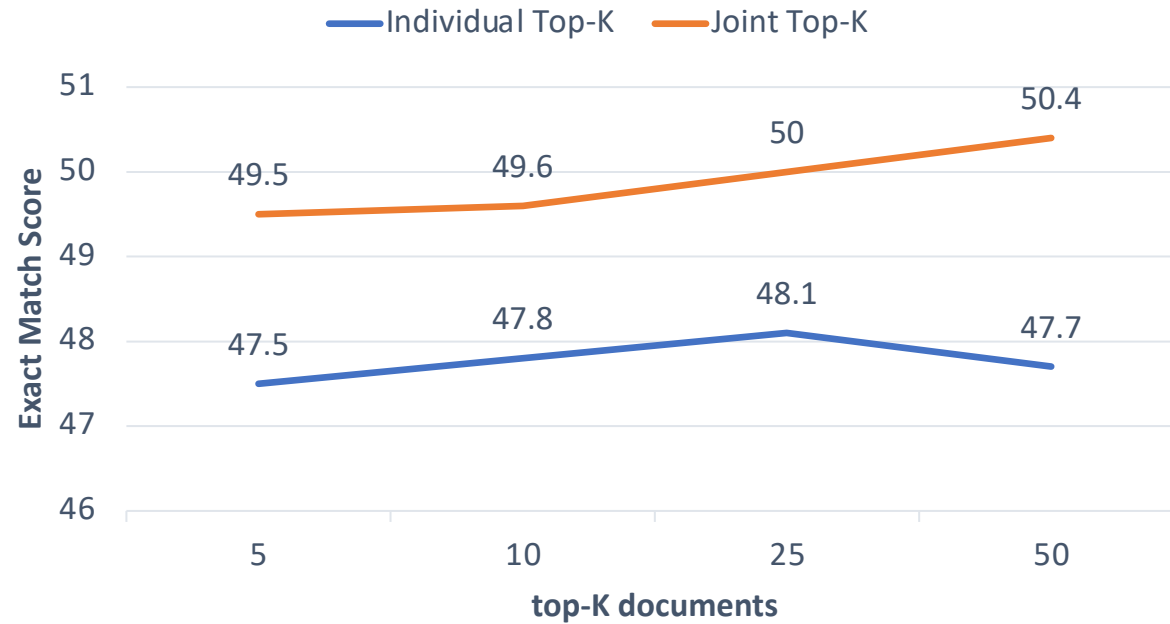
Competitive performance with FiD

Effect of Retriever Update on NQ Dev



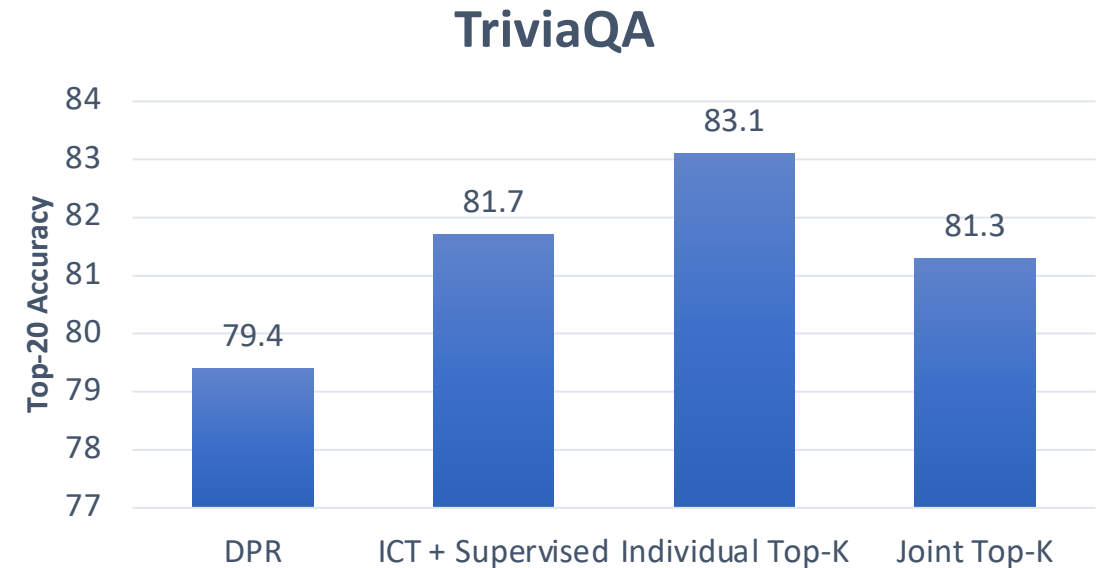
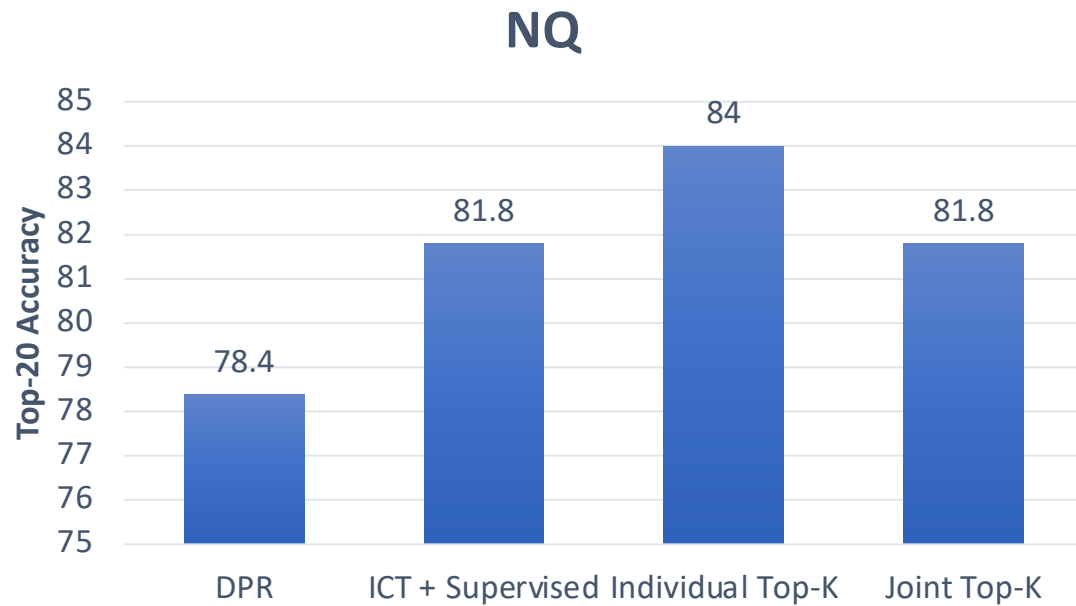
Retriever score bias helps for smaller top-K values

End-to-End Approaches Comparison: Answer Extraction



1. Joint Top-K is more effective
2. Performance increases with more top-k documents

End-to-End Approaches Comparison: Retrieval Accuracy



Individual Top-k training is more effective in improving retrieval.

Key Takeaways

1. Retriever score scaling helps during training.
2. Unsupervised ICT and MSS training **improves** retrieval performance.
3. **Joint inter-attention** of top-K documents **outperforms** individual encoding for answer extraction.
4. **Biasing** inter-attention with **retriever score** improves answer extraction for smaller top-k values.

Thank You!

- Paper: <https://arxiv.org/abs/2101.00408>
- Code: <https://github.com/NVIDIA/Megatron-LM/tree/main/tasks/orqa>
- Contact:
 - Devendra Sachan (sachande@mila.quebec)
 - Mostofa Patwary (mpatwary@nvidia.com)

Approaches which didn't work

- DPR with RoBERTa didn't give much improvements
- ICT with batch size of 8K diverged in training
- Training ICT for more than 100K steps didn't give improvements.
- MSS trained for more iterations didn't give improvements.

- Major bug which still worked for DPR
 - Flipping the attention masks
 - Attending to just the padded tokens