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

What Nvidia card does the surface book have?

GeForce GTX 965M

When is season seven of Game of Thrones coming out?

July 16 , 2017

Five Australian prime ministers who governed during World War 2

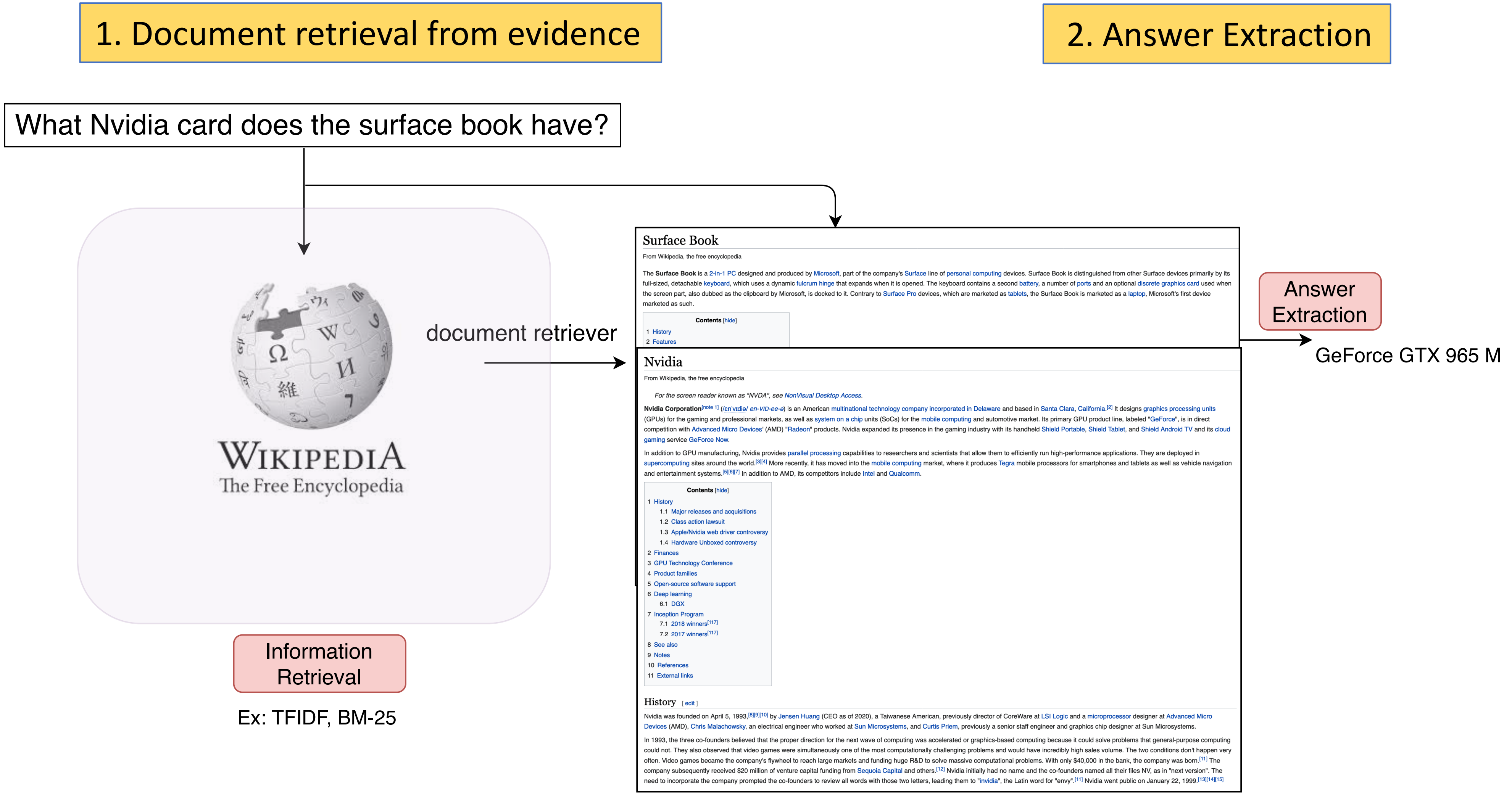
Arthur Fadden, John Curtin, Frank Forde, Robert Menzies, Ben Chifley

WIKIPEDIA
The Free Encyclopedia

Chen et al., 2017. Reading Wikipedia to Answer Open-domain Questions

Background: Open-Domain QA

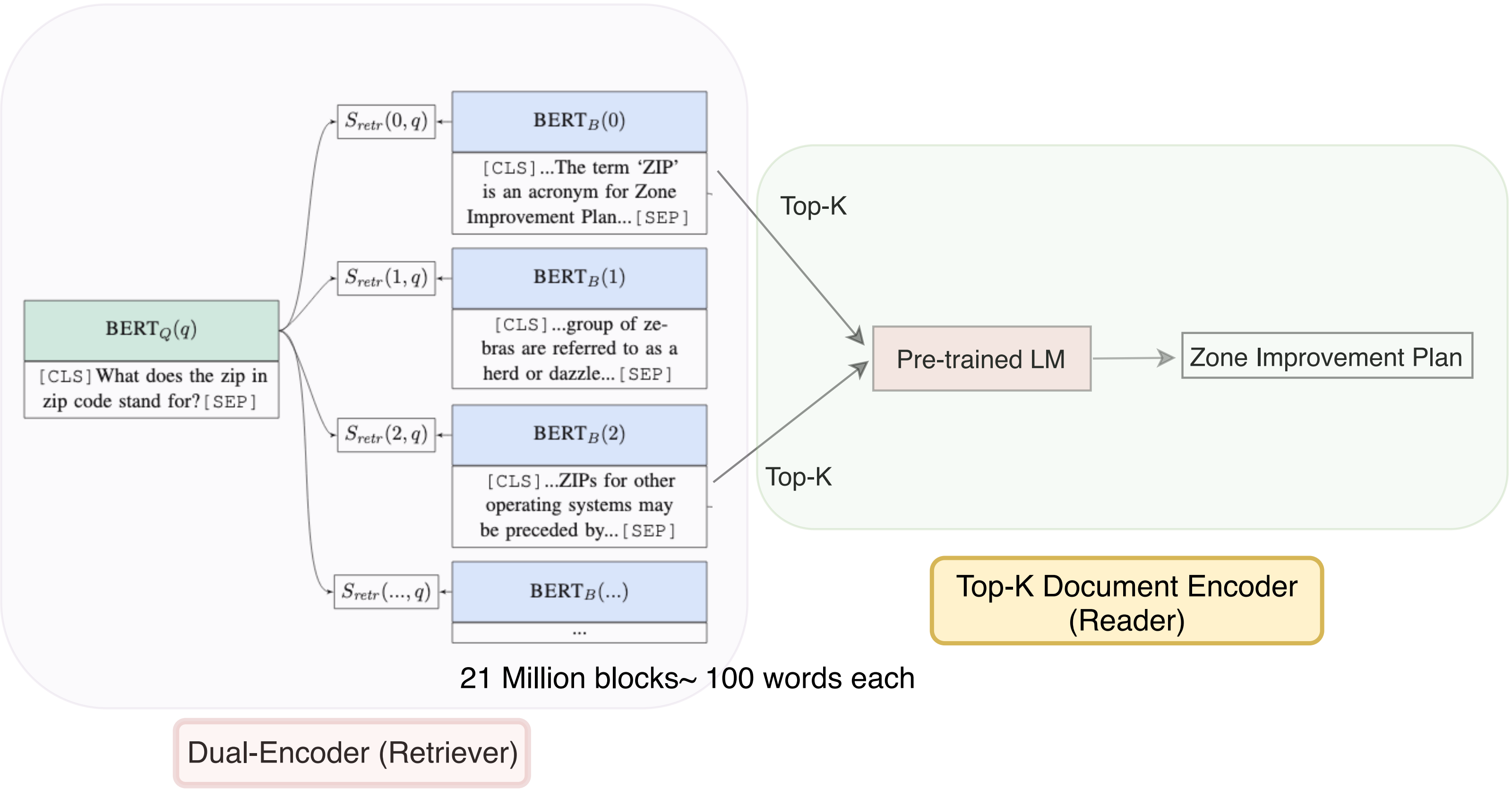
- Two-stage approach



Background: Neural Models for Open-Domain QA

Learned Information Retrieval

Learned Answer Extraction

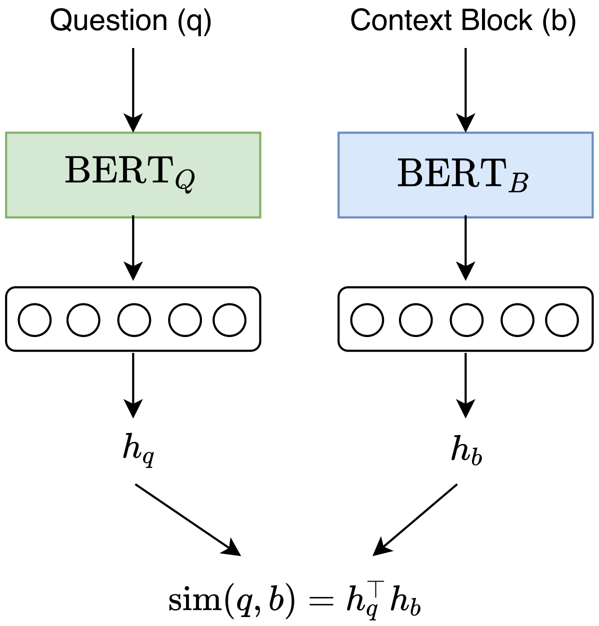


Previous Work
1. ORQA: Lee et al. 2019
2. REALM: Guo et al. 2020

Image from "Lee et al., 2019. Latent Retrieval for Weakly Supervised Open Domain Question Answering"

Prior Work: Learned Information Retrieval

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



$$D = q_i,$$ Query

$$b_i^+,$$ Positive Context

$$b_j^-$$ Other Context

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

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.

Karpukhin et al., 2020. Dense Passage Retrieval for Open-Domain Question Answering

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*.

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

  - 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.

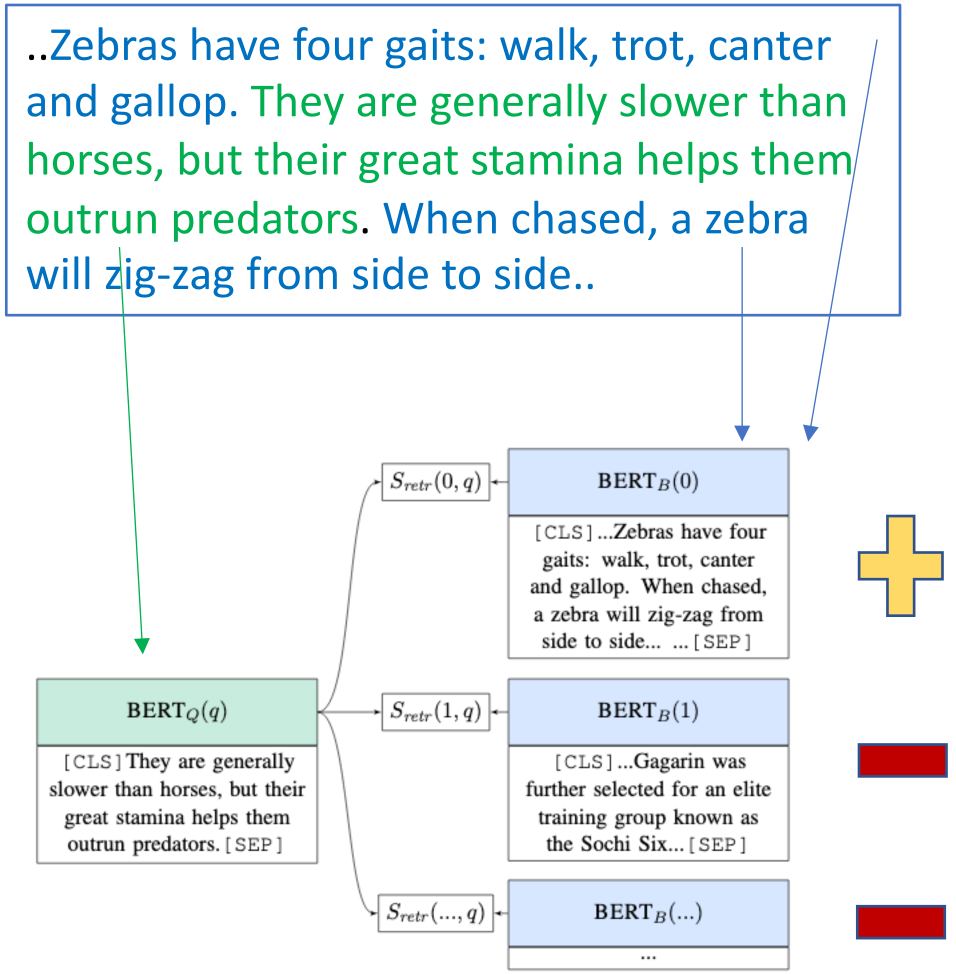


Image from "Lee et al., 2019. Latent Retrieval for Weakly Supervised Open Domain Question Answering"

Retriever Pre-training by Masked Salient Spans

- Masked Salient Spans (MSS) training: first proposed in Guu at 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.

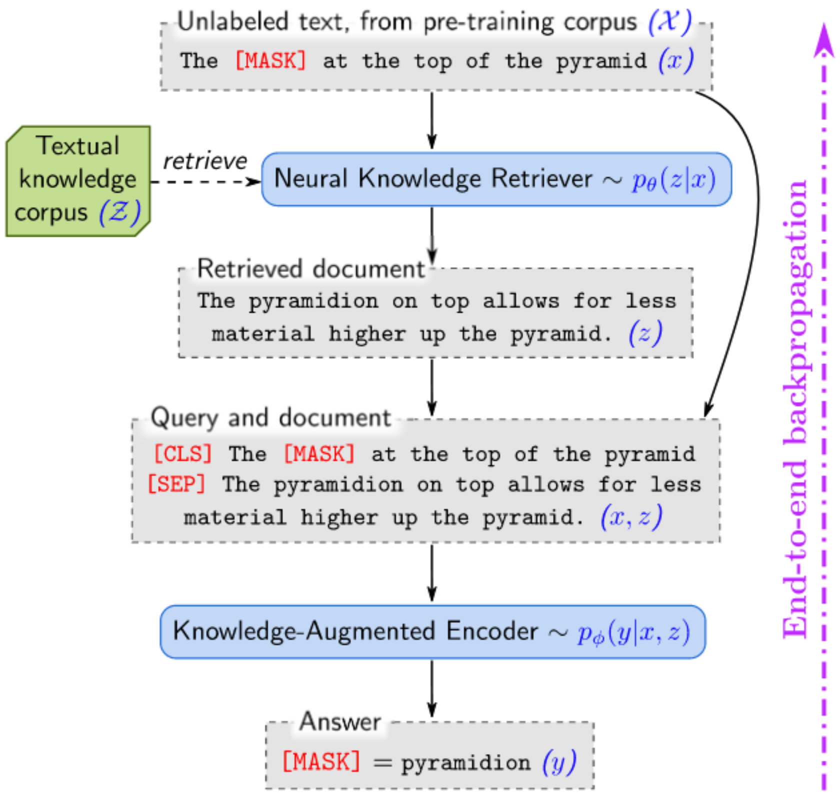


Image from "Guu et al., 2020. REALM: Retrieval-Augmented Language Model Pre-Training "

Experimental Setup

Datasets

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

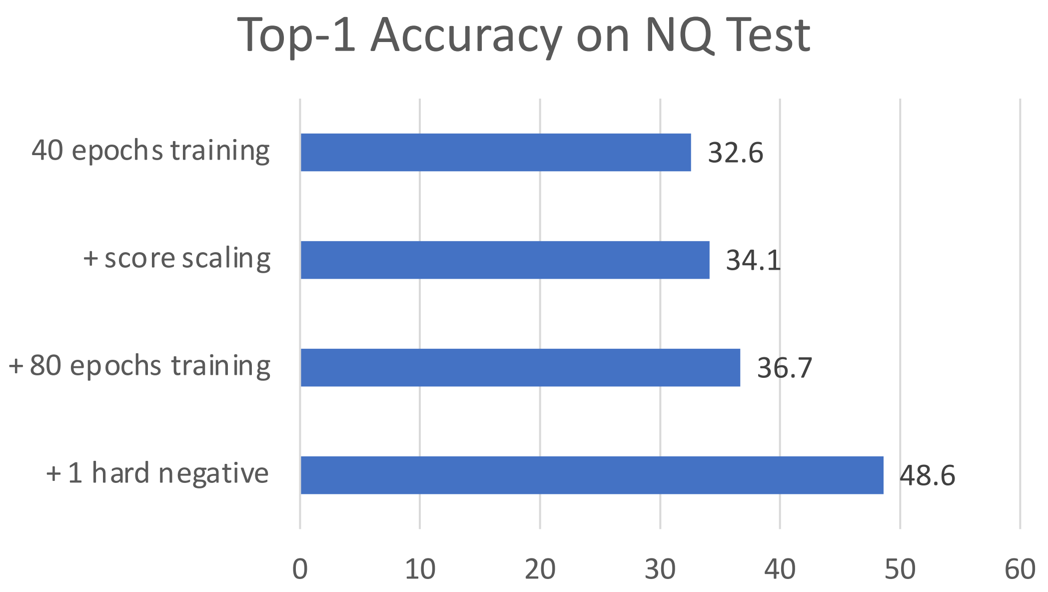
Dataset	Train	Val	Test
NQ	79,168	8,757	3,610
TriviaQA	78,785	8,837	11,313

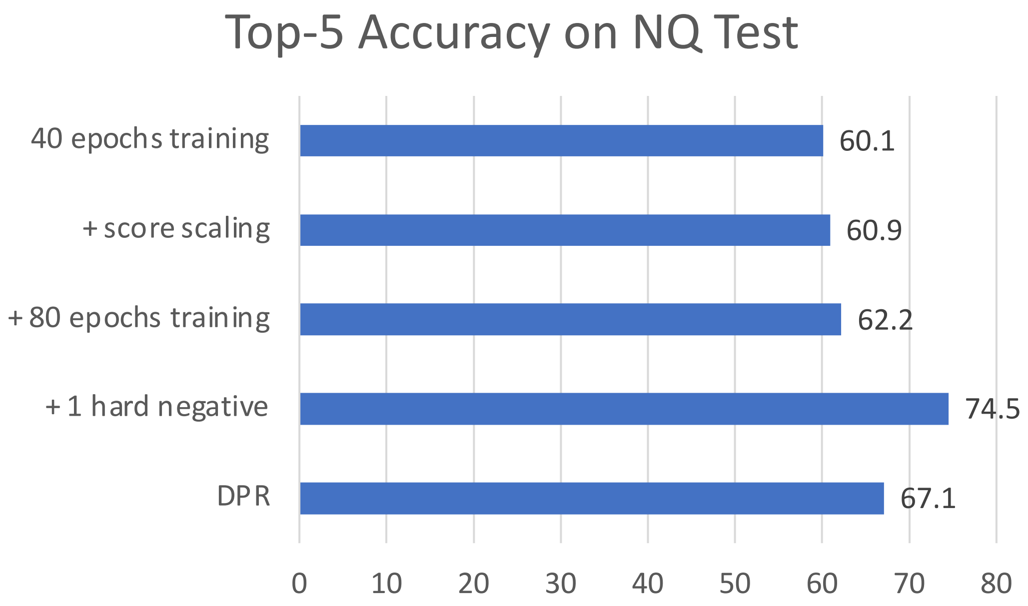
Evaluation Metric

Precision@top-K

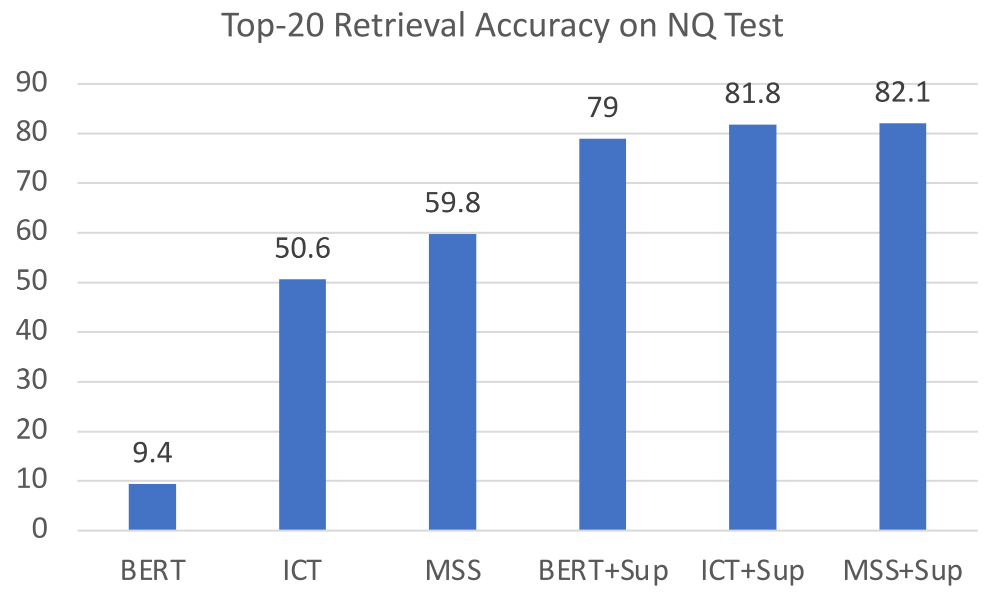
- Exact Match if answers exists in top-K documents or not

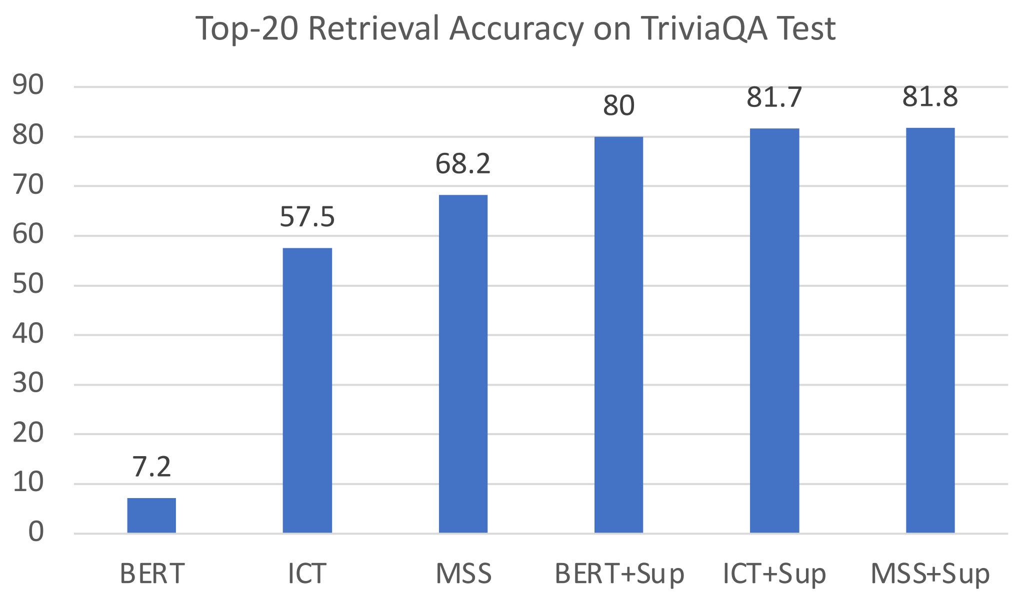
Results: Effect of Score Scaling and Longer Training





Results: Effect of Unsupervised Pre-training

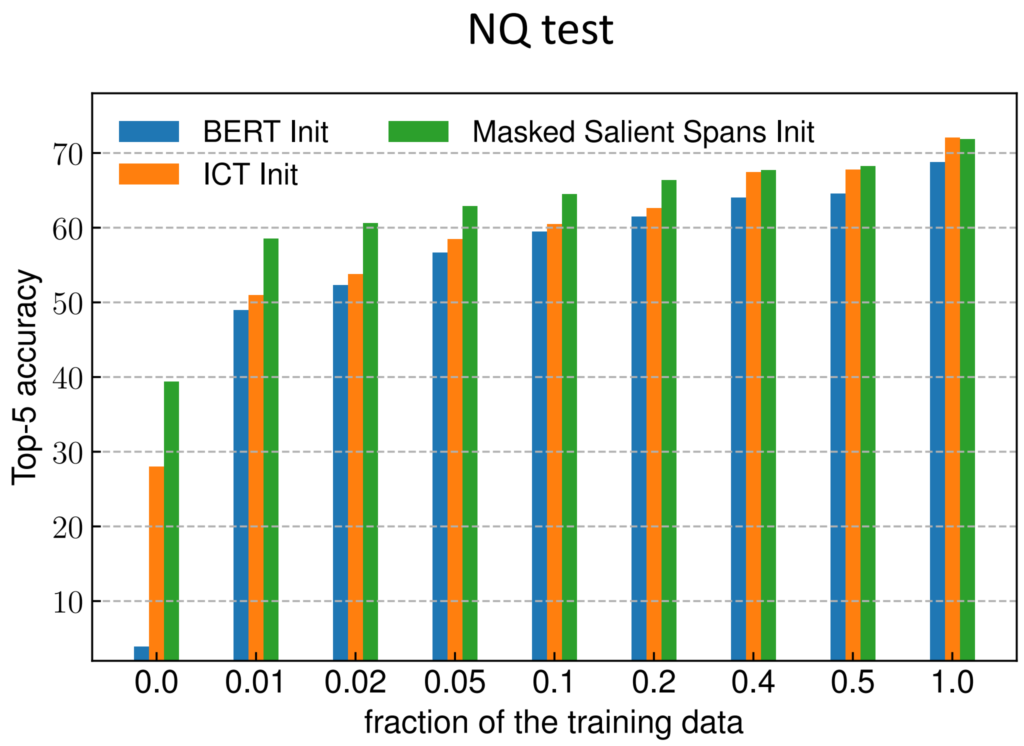


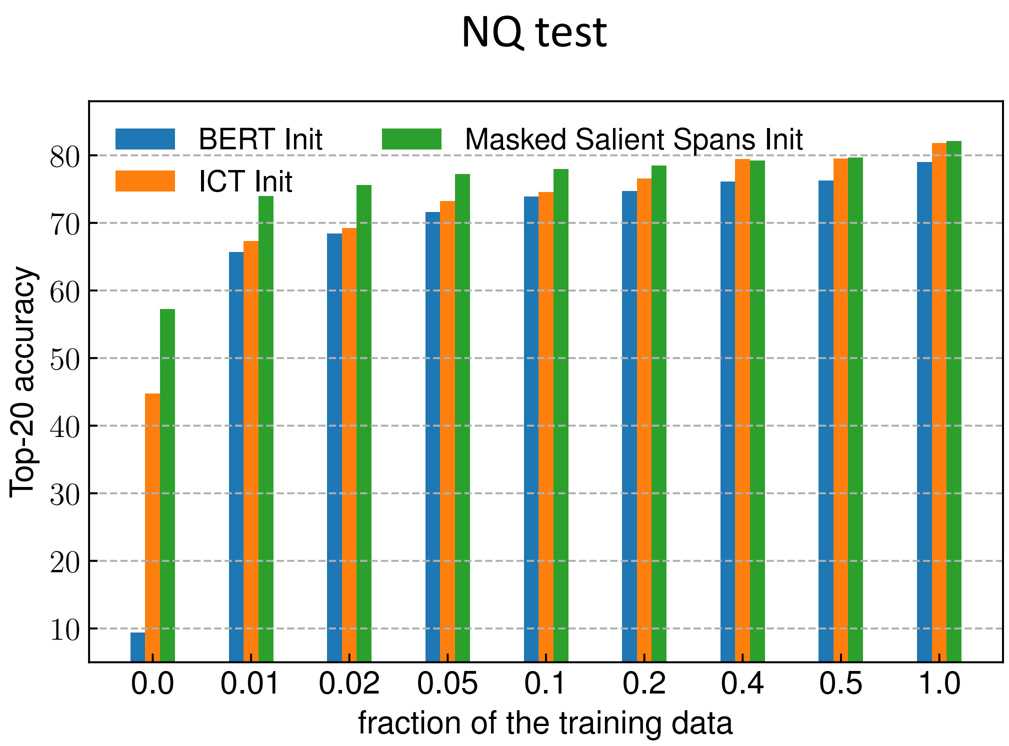


ICT + Supervised and MSS + Supervised outperform Supervised retriever training

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

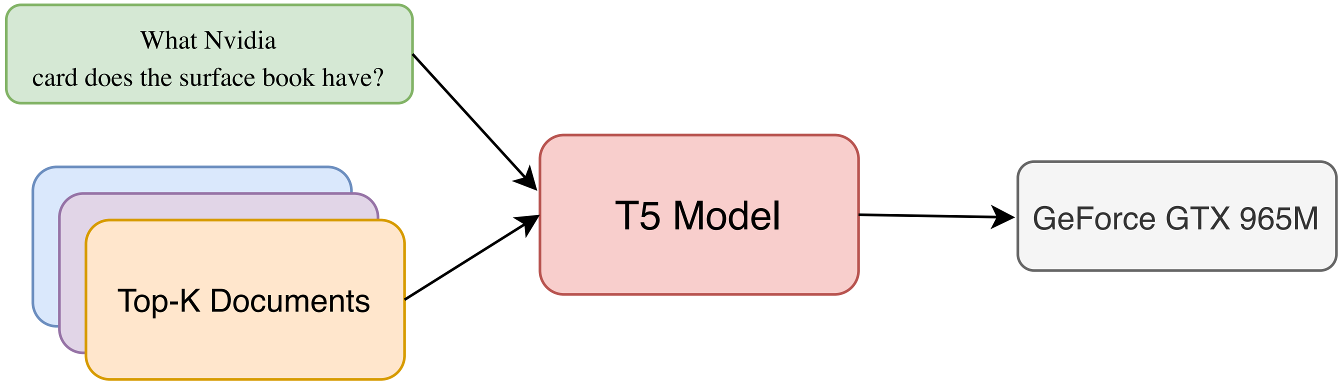
End-to-End Supervised Training

Neural Reader for Answer Extraction

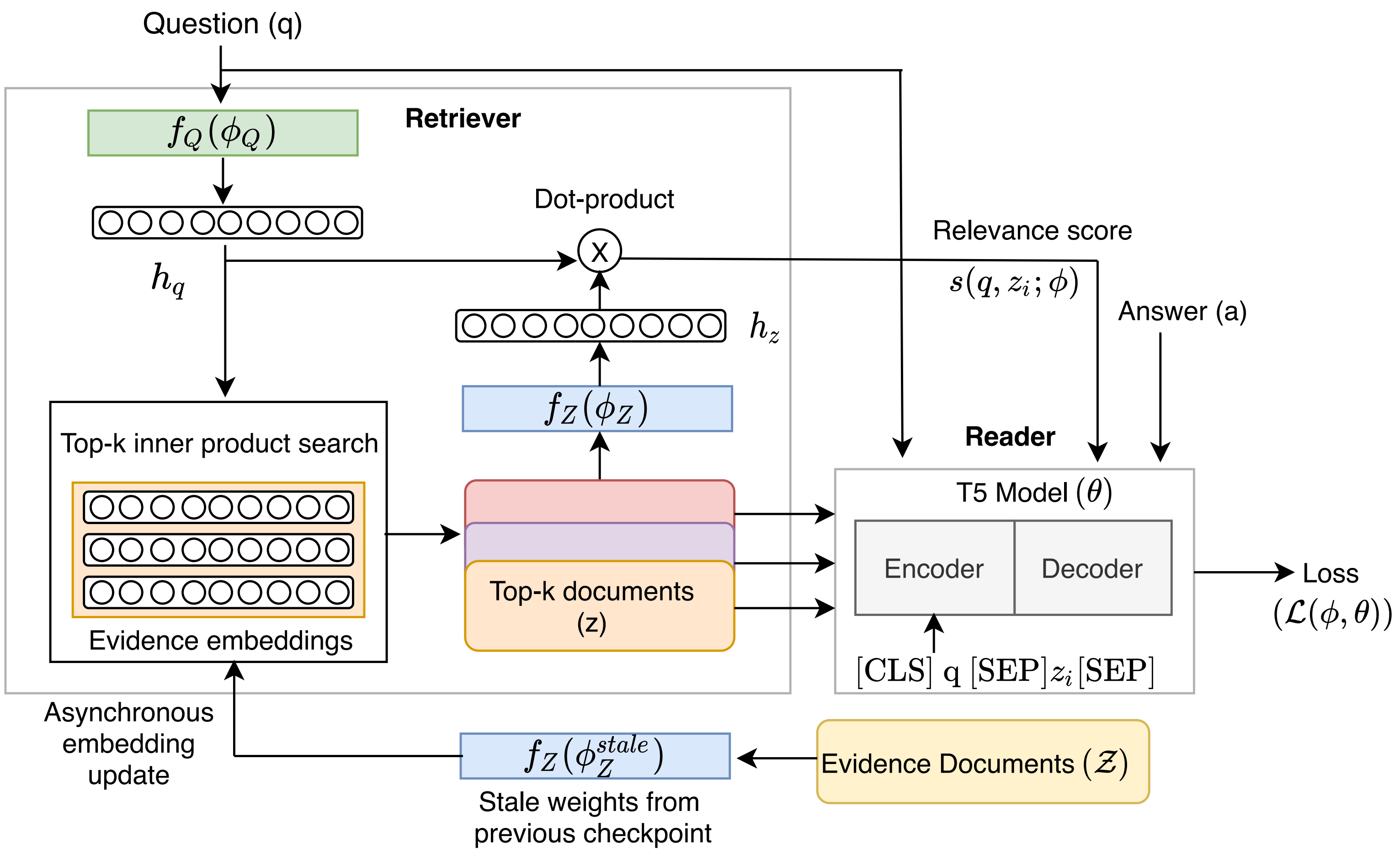
1. Retrieve top-K documents from **retriever**

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

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 = *question*, a = *answer*, z = *top-K doc*

Computed from T5

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

Computed from Retriever

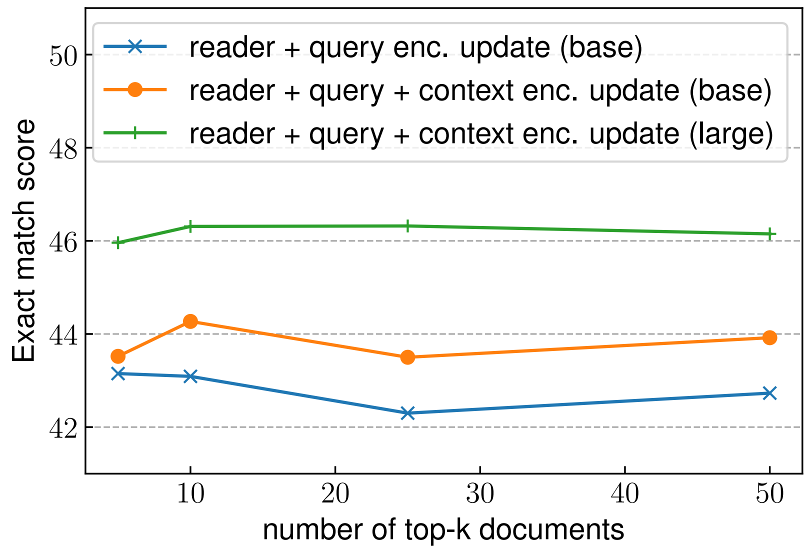
Decoding: Select the best answer among K answers

Results: Individual Top-K

Model	NQ	TriviaQA
Base Configuration		
ORQA	33.3	45.0
REALM	40.4	–
DPR	41.5	56.8
Individual Top-k	**45.9**	56.3
Large Configuration		
RAG	44.5	56.8
Individual Top-k	**48.1**	**59.6**

Better Retriever + T5 top-K encoder helps!

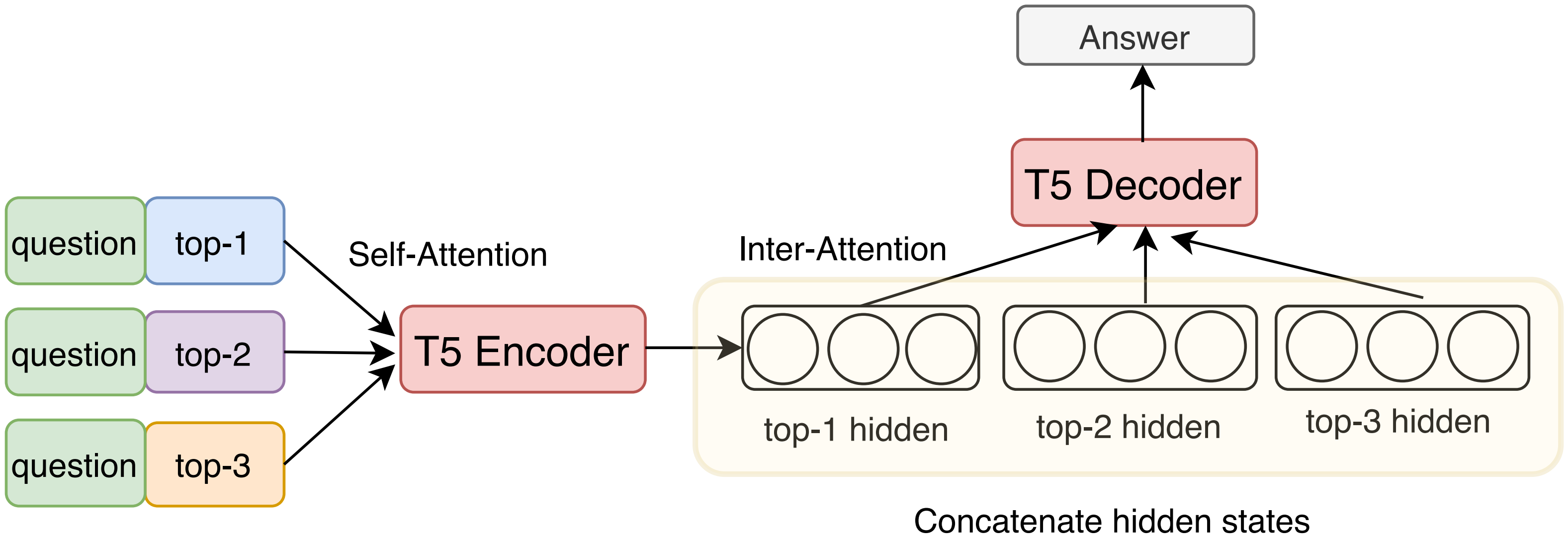
Effect of Async Retriever Update on NQ Dev



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



$$\mathrm{attention}(q, a) \propto Q(a)K(x_1 \ldots x_k; q) + \beta p(x_i \mid q)$$

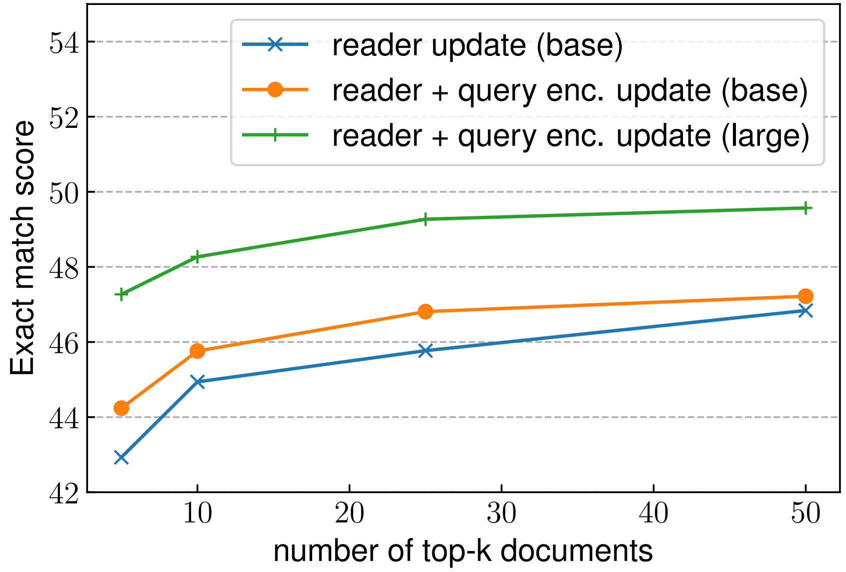
Retriever similarity score bias

Results: Joint Top-K

Model	NQ	TriviaQA
Base Configuration		
FiD	48.2	65.0
Joint Top-k	**49.2**	64.8
Large Configuration		
FiD	51.4	67.6
Joint Top-k	**51.4**	**68.3**

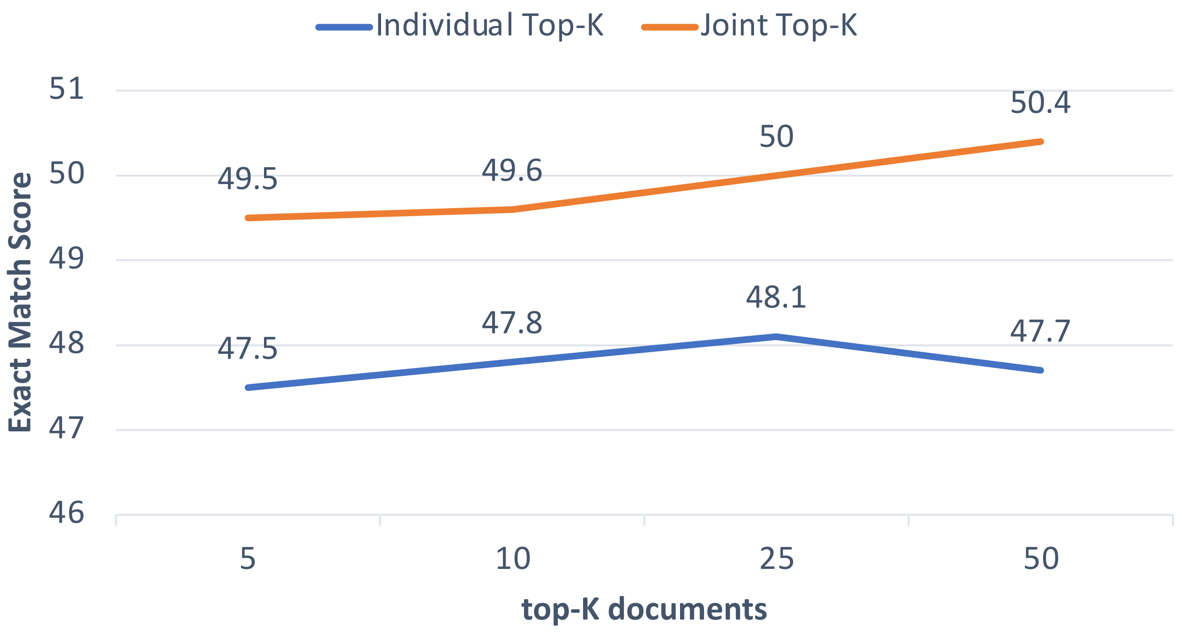
Competitive performance with FiD

Effect of Retriever Update on NQ Dev



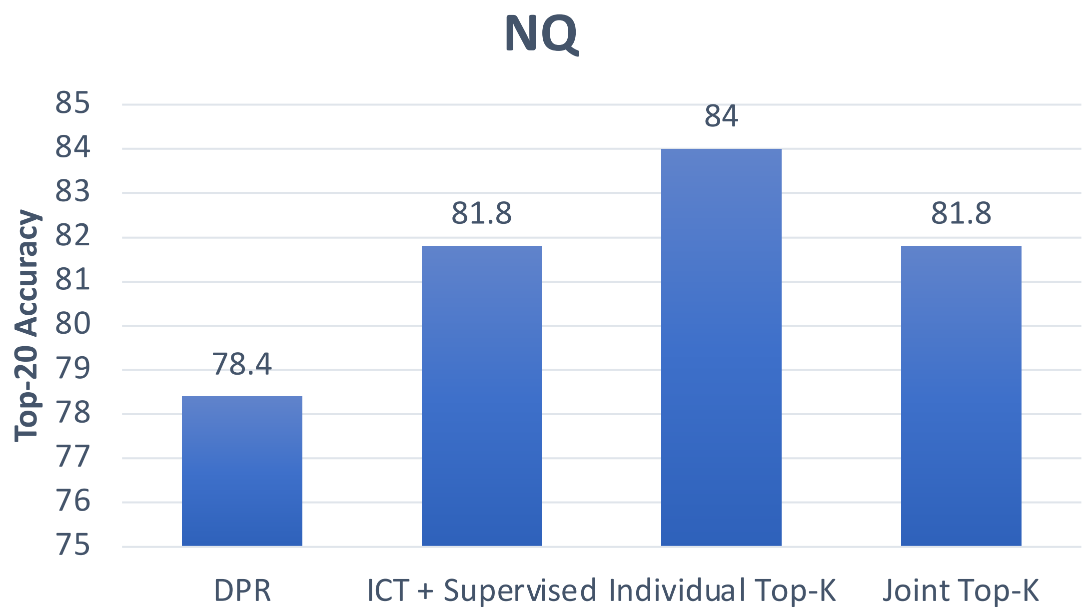
Retriever score bias helps for smaller top-K values

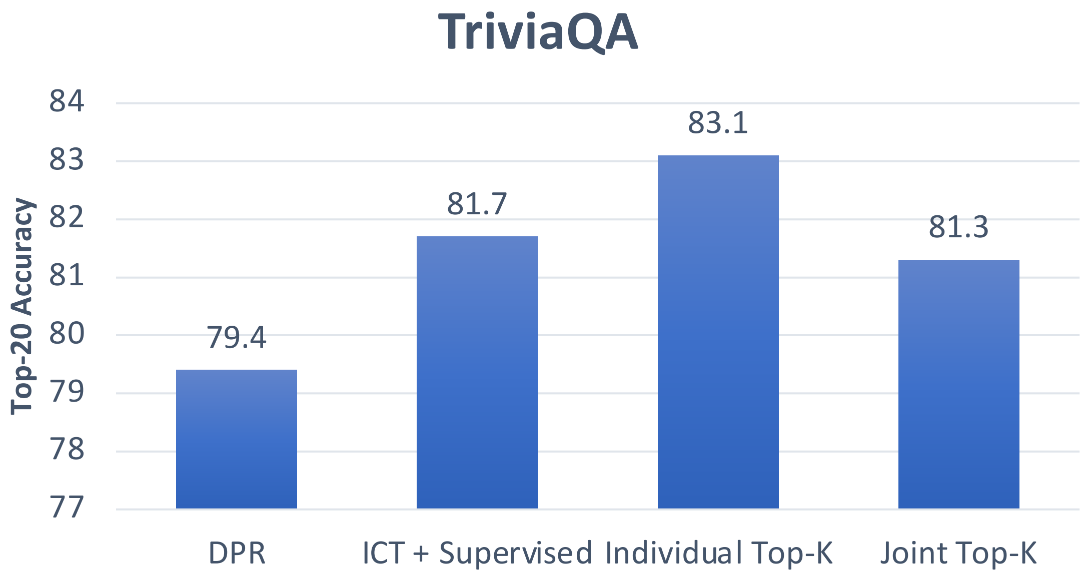
End-to-End Approaches Comparison: Answer Extraction



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