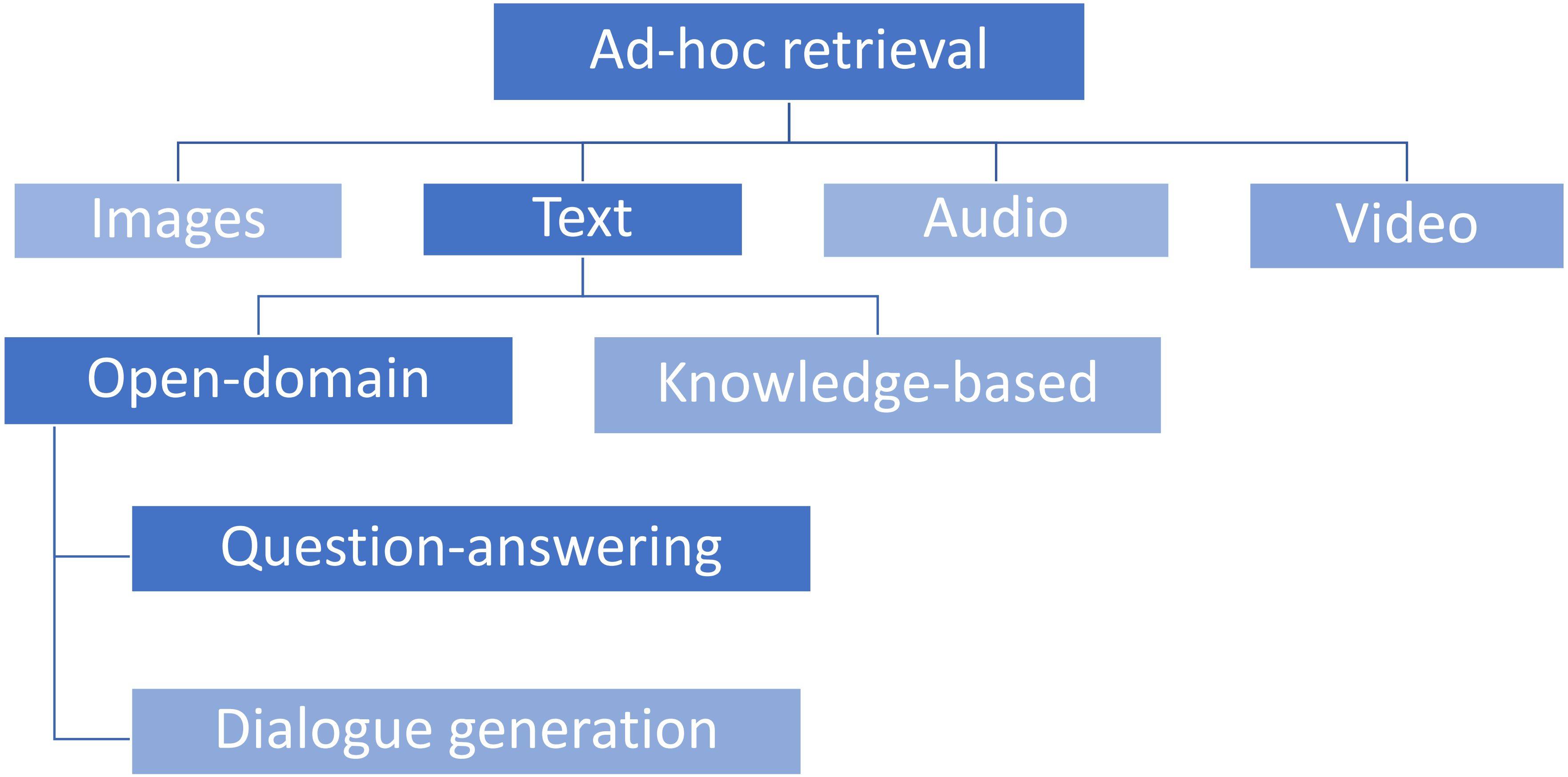
Zero-Shot Approach to Train Dense Retriever for Question Answering

- Devendra Singh Sachan

Background: Terminology



Background: Information-Seeking (factoid) Questions

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

What Nvidia card does the surface book have? → GeForce GTX 965M

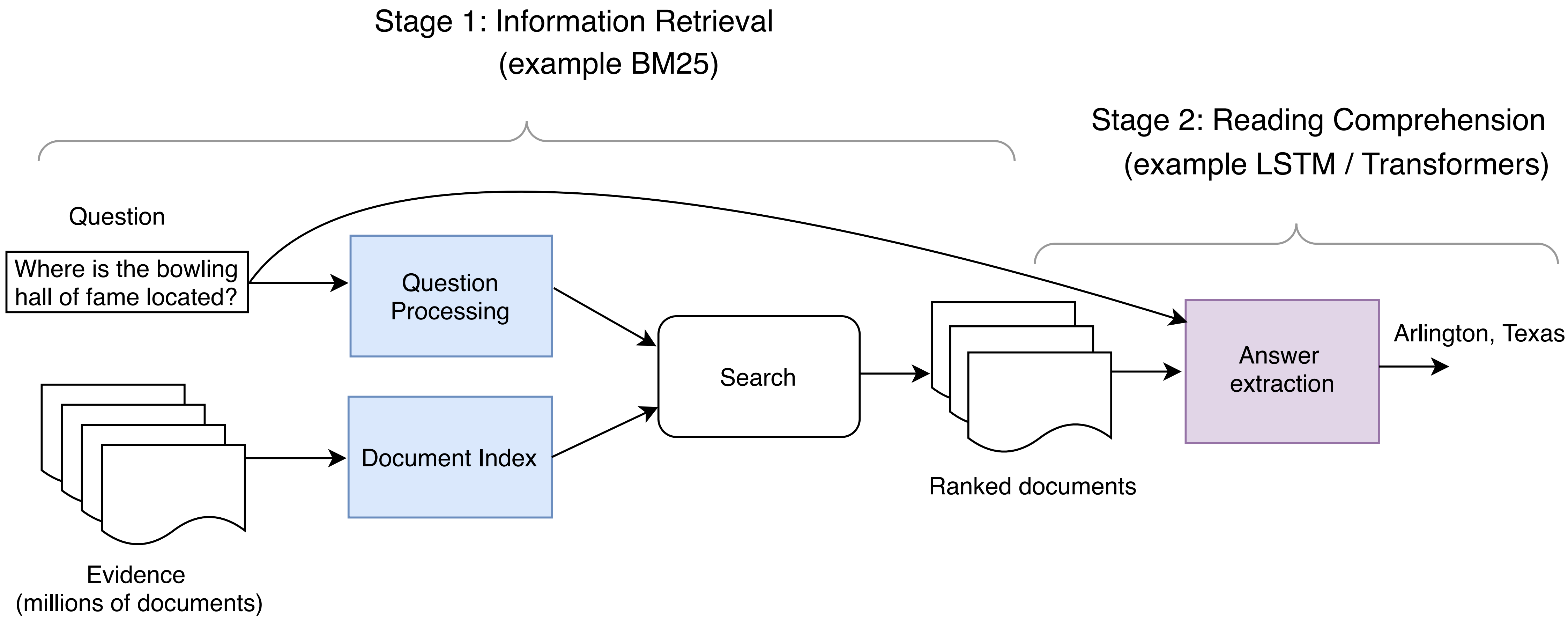
When is season seven of Game of Thrones coming out? → July 16 , 2017

Where is the bowling hall of fame located? → Arlington, Texas

WIKIPEDIA
The Free Encyclopedia

Evidence documents

Open-Domain QA Pipeline (pre-2018)



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

Strong Baseline: BM25

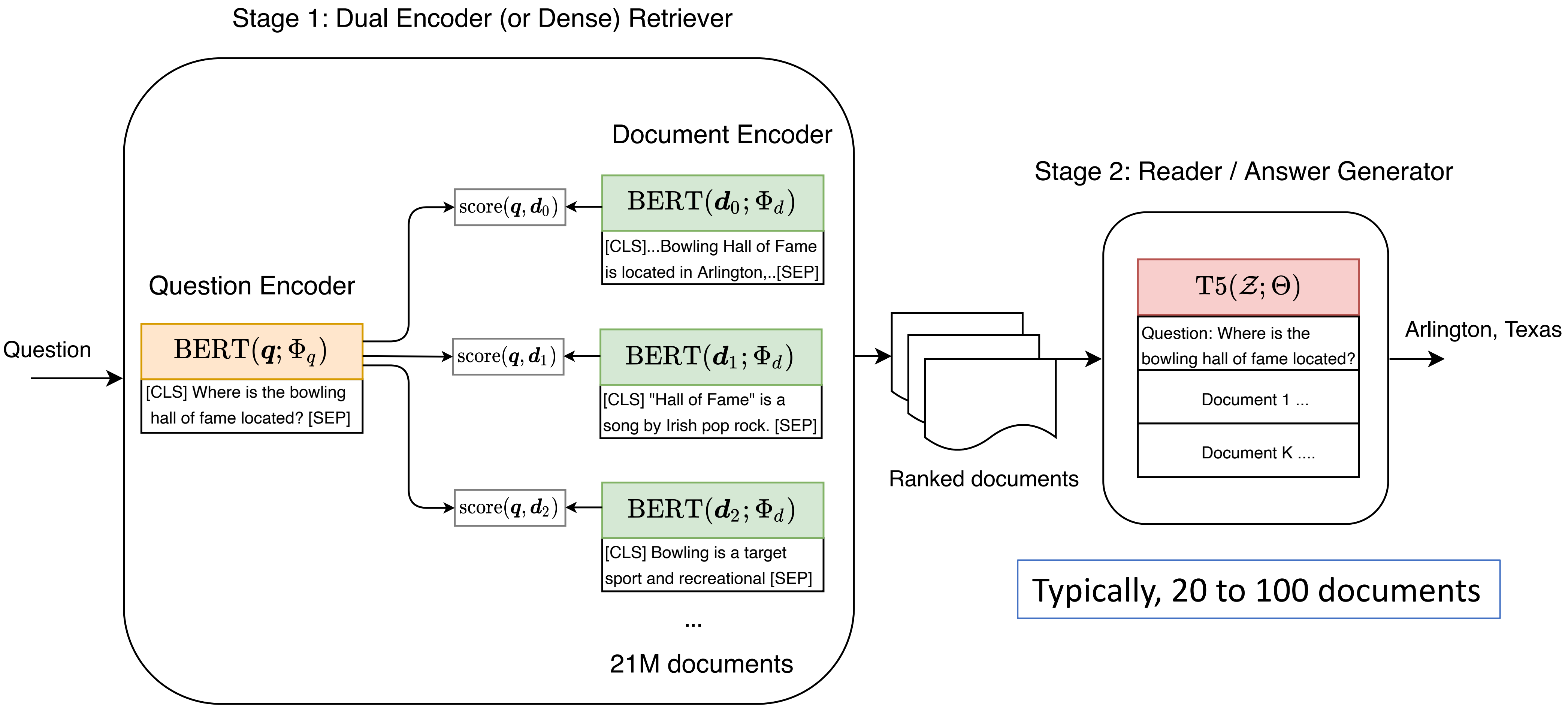
- Sparse vector-space approach based on bag-of-words assumption

Given a query Q, containing keywords $q_1, \ldots, q_n$, the BM25 score of a document D is:

$$\text{score}(D,Q) = \sum_{i=1}^{n} \text{IDF}(q_i) \cdot \frac{f(q_i, D) \cdot (k_1 + 1)}{f(q_i, D) + k_1 \cdot \left(1 - b + b \cdot \frac{|D|}{\text{avgdl}}\right)}$$

- k_1 and b are hyper-parameters.
- Popular implementations: Gensim, Lucene

Neural Models for Open-Domain QA (post-2018)

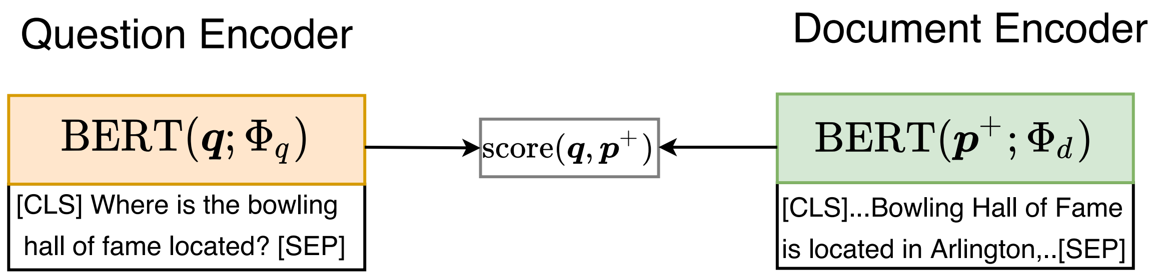


Typically, 20 to 100 documents

Latent Retrieval for Weakly Supervised Open Domain Question Answering, Lee et al., ACL 2019 .

Review: DPR Training

Task: Train dual-encoder to improve retrieval accuracy



Training data: *Question, positive, negative documents*

$$\mathcal{D} = \{\langle q_i, p_i^+, p_{i,1}^-, \dots, p_{i,n}^- \rangle\}_{i=1}^{\mathcal{T}}$$

$$\text{score}(\boldsymbol{q}, \boldsymbol{d}_i; \Phi) = f_q(\boldsymbol{q}; \Phi_q)^\top f_d(\boldsymbol{d}_i; \Phi_d)$$

Contrastive Training Objective

$$L = -\log \frac{e^{\text{score}(q_i, p_i^+)}}{e^{\text{score}(q_i, p_i^+)} + \sum_{j=1}^{n} e^{\text{score}(q_i, p_{ij}^-)}}$$

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

DPR Training: Impact

Training	Retriever	Top-20 NQ	Top-20 TriviaQA	Top-20 WQ	Top-20 TREC	Top-20 SQuAD	Top-100 NQ	Top-100 TriviaQA	Top-100 WQ	Top-100 TREC	Top-100 SQuAD
None	BM25	59.1	66.9	55.0	70.9	68.8	73.7	76.7	71.1	84.1	80.0
Single	DPR	78.4	79.4	73.2	79.8	63.2	85.4	**85.0**	81.4	89.1	77.2
	BM25 + DPR	76.6	79.8	71.0	85.2	**71.5**	83.8	84.5	80.5	92.7	**81.3**
Multi	DPR	**79.4**	78.8	**75.0**	**89.1**	51.6	**86.0**	84.7	**82.9**	93.9	67.6
	BM25 + DPR	78.0	**79.9**	74.7	88.5	66.2	83.9	84.4	82.3	**94.1**	78.6

- DPR obtains 10-20 points improvement over BM25 on multiple benchmarks
- This has led to good end-task performance such as QA
- Widely used since it was introduced (> *700 citations in last 2 years*)

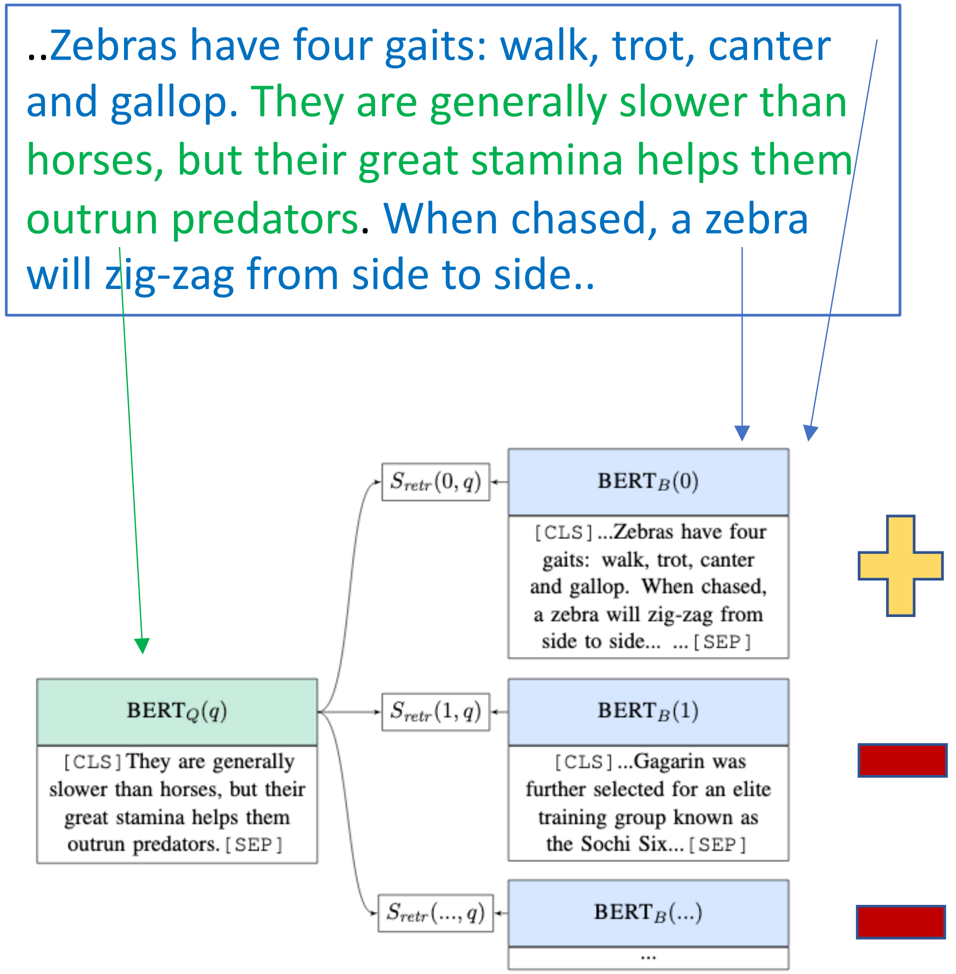
Dense Passage Retrieval for Open-Domain Question Answering, Karpukhin et al., EMNLP 2020.

DPR Training: Limitations

1. Require aligned documents for training
   - Expensive to annotate thousands of positive documents for peak performance
2. Hard-negative documents: dependent **on BM25 outputs**
   - Needs to be pre-computed
3. Expensive GPU communication operations in forward pass when scaling up batch size

Review: Unsupervised (Dense) Approaches

- 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.
- Examples: **ICT, Contriever**



ICT: Latent Retrieval for Weakly Supervised Open Domain Question Answering, Lee et al., ACL 2019
Contriever: Unsupervised Dense Information Retrieval with Contrastive Learning, Izacard et al., TMLR 2022

Summary: Unsupervised (Dense) Approaches

- BM25 still a stronger baseline than dense unsupervised methods.
- **Large performance gap** when compared with supervised approaches.
- **Research Question**: How can we improve unsupervised retrievers with minimal supervision?

Questions Are All You Need to Train a Dense Passage Retriever

Authors: Devendra Singh Sachan, Mike Lewis, Dani Yogatama, Luke Zettlemoyer, Joelle Pineau, and Manzil Zaheer

Mike Lewis

Dani Yogatama

Luke Zettlemoyer

Joelle Pineau

Manzil Zaheer

Autoencoding-based Retriever Training (ART)

- **Training Data**: Uses only questions and evidence documents

$$\mathcal{D} = \{q_i\}_{i=1}^{\mathcal{T}}$$

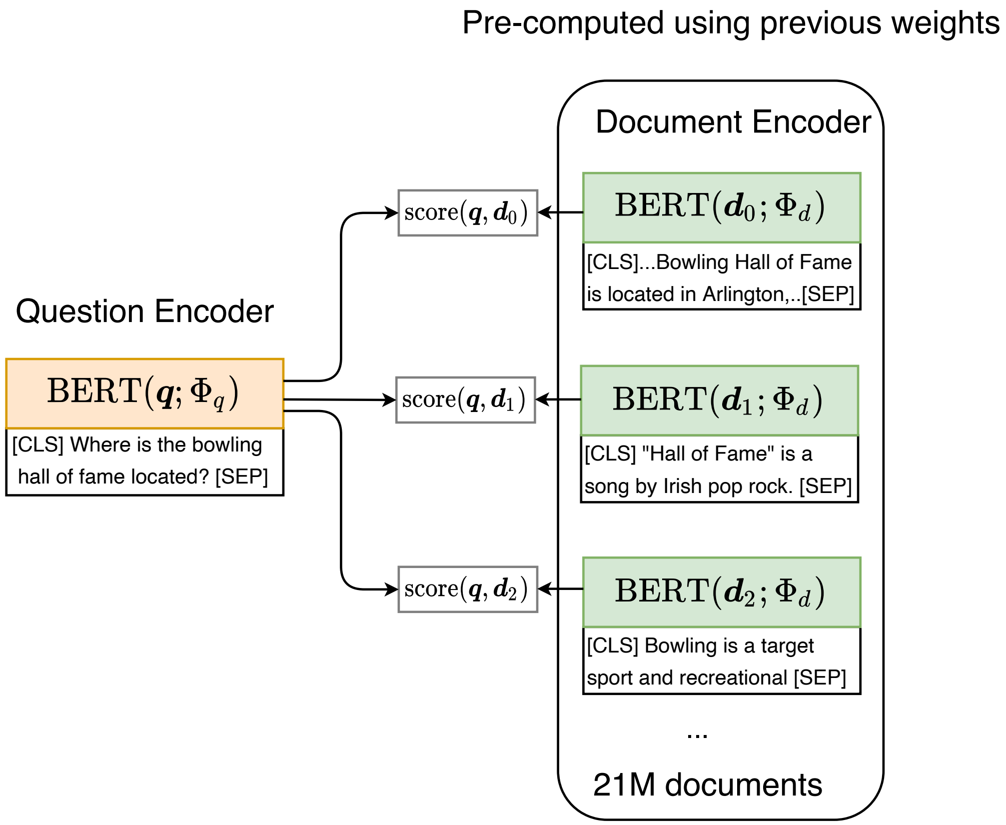
- **Auto-encoder Intuition**: Retrieve using and generate the same question

	DPR	Our Method
Zero-shot	✖	✓
BM25 negatives	✓	✖
Simplified training	✖	✓

We address DPR training limitations in our work and just use only questions

ART: Training Details

Step 1a. Compute question similarity with all evidence documents



Evidence Documents

$$\mathcal{D} = \{\boldsymbol{d}_1, \ldots, \boldsymbol{d}_M\}$$

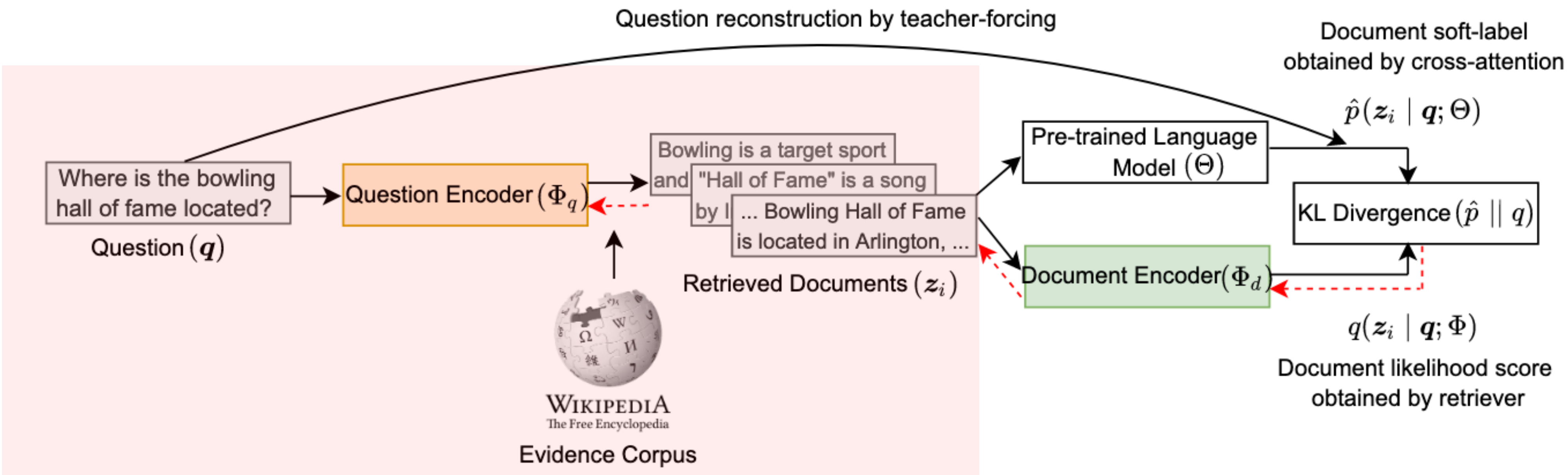
$$\text{score}(\boldsymbol{q}, \boldsymbol{d}_i; \Phi) = f_q(\boldsymbol{q}; \Phi_q)^\top f_d(\boldsymbol{d}_i; \Phi_d)$$

- Dot-product is highly optimized on GPUs

ART: Training Details

Step 1b. Select top-K documents with highest scores

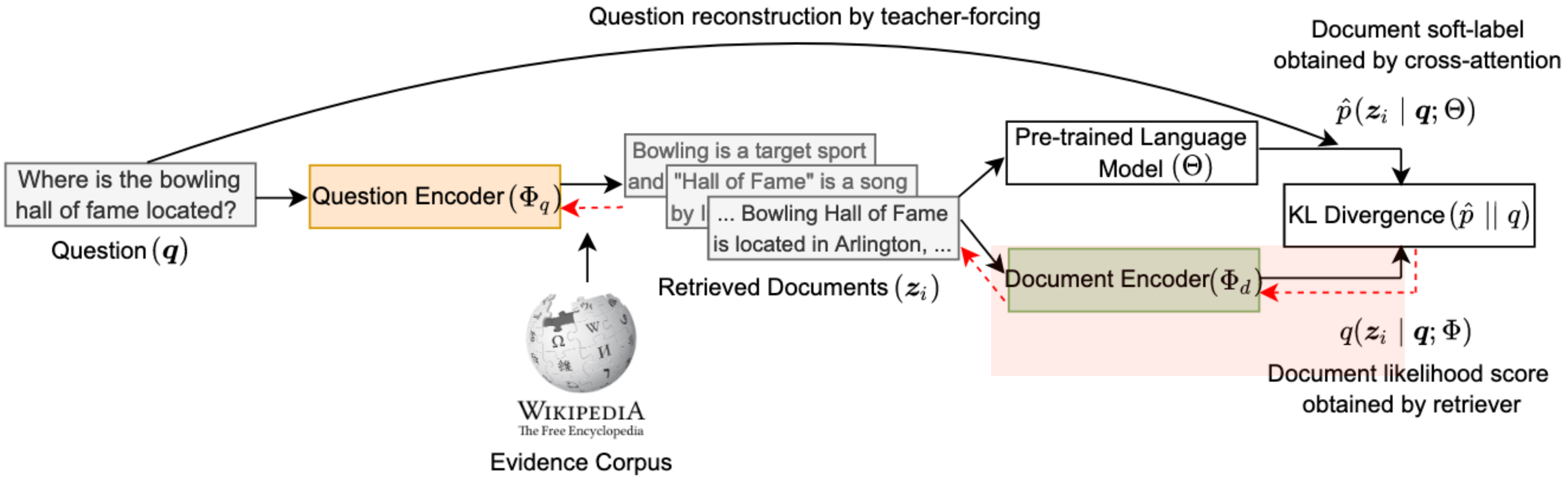
$$\mathcal{Z} = \{z_1, \ldots, z_K\}$$



ART: Training Details

Step 2a. Calculate scores using "*current*" document encoder weights

$$\text{score}(\boldsymbol{q}, \boldsymbol{z}_i; \Phi) = f_q(\boldsymbol{q}; \Phi_q)^\top f_d(\boldsymbol{z}_i; \Phi_d)$$

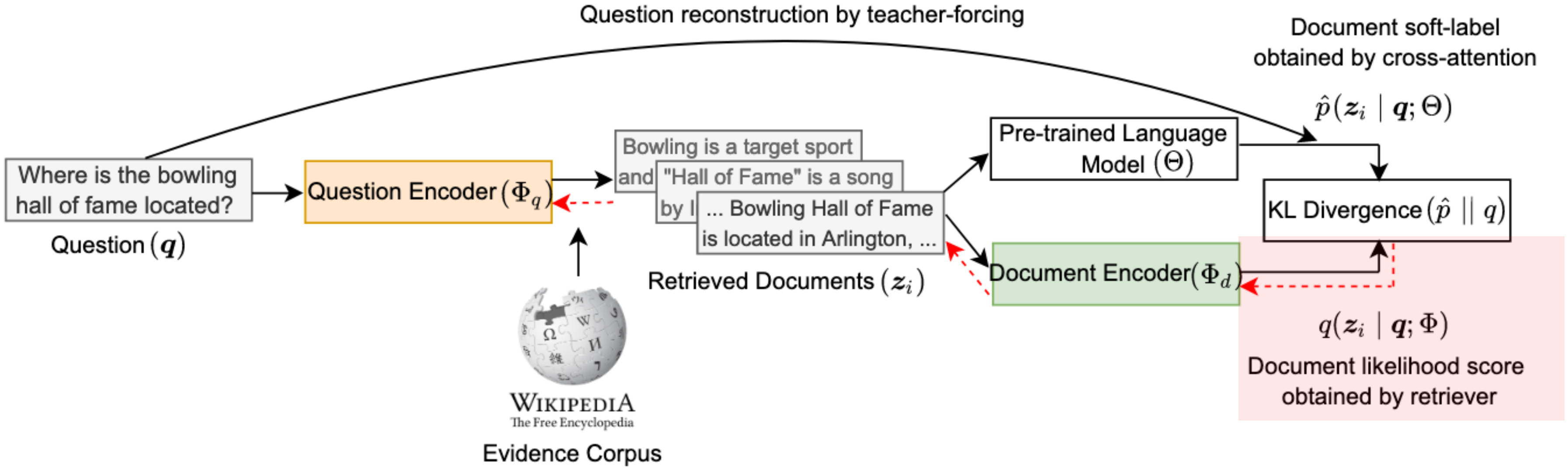


ART: Training Details

Step 2b. Calculate retriever distribution

$$q(\boldsymbol{z}_i \mid \boldsymbol{q}, \mathcal{Z}; \Phi) \propto \text{score}(\boldsymbol{q}, \boldsymbol{z}_i)$$

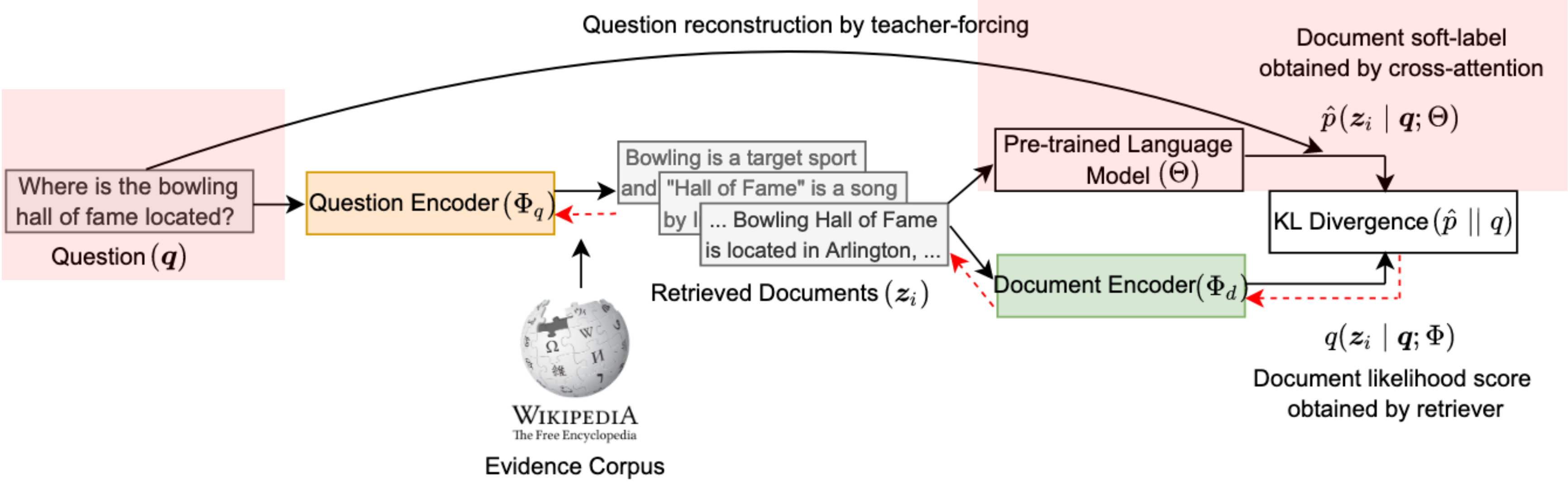
$$q(\boldsymbol{z}_i \mid \boldsymbol{q}, \mathcal{Z}; \Phi) = \frac{\exp(\text{score}(\boldsymbol{q}, \boldsymbol{z}_i)/\tau)}{\sum_{j=1}^{K} \exp(\text{score}(\boldsymbol{q}, \boldsymbol{z}_j)/\tau)}$$



ART: Training Details

Step 3a. PLM relevance score calculation by question reconstruction

$$\log p(\boldsymbol{z}_i \mid \boldsymbol{q}; \Theta) \propto \frac{1}{|\boldsymbol{q}|} \sum_t \log p(q_t \mid \boldsymbol{q}_{<t}, \boldsymbol{z}_i; \Theta)$$



PLM Relevance Score: Details

PLM relevance score calculation by question reconstruction

$$\log p(\boldsymbol{z}_i \mid \boldsymbol{q}; \Theta) \propto \frac{1}{|\boldsymbol{q}|} \sum_t \log p(q_t \mid \boldsymbol{q}_{<t}, \boldsymbol{z}_i; \Theta)$$

- **Accurate**: cross-attention between question and document
- **Unsupervised**: just perform inference using PLM
- Choice of PLM is important, we use **T0-3B**

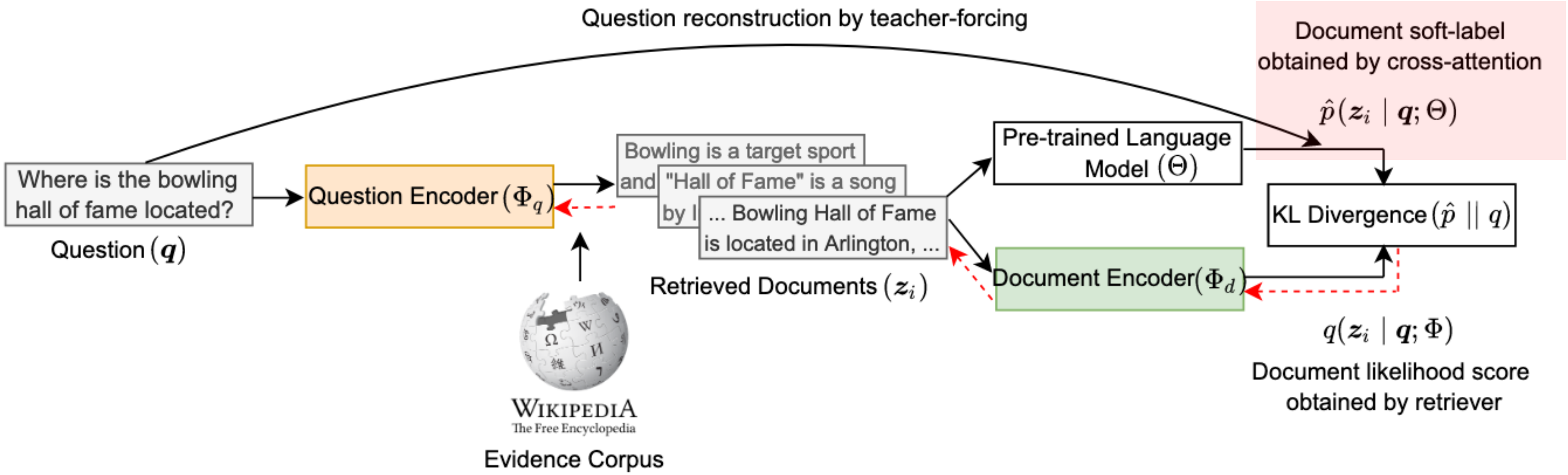
T0-3B: Multitask prompted training enables zero-shot task generalization , Sanh et al., ICLR 2022.

ART: Training Details

Step 3b. Calculate distribution over PLM scores

$$\hat{p}(\boldsymbol{z}_i \mid \boldsymbol{q}, \mathcal{Z}) \propto \log p(\boldsymbol{z}_i \mid \boldsymbol{q}; \Theta)$$

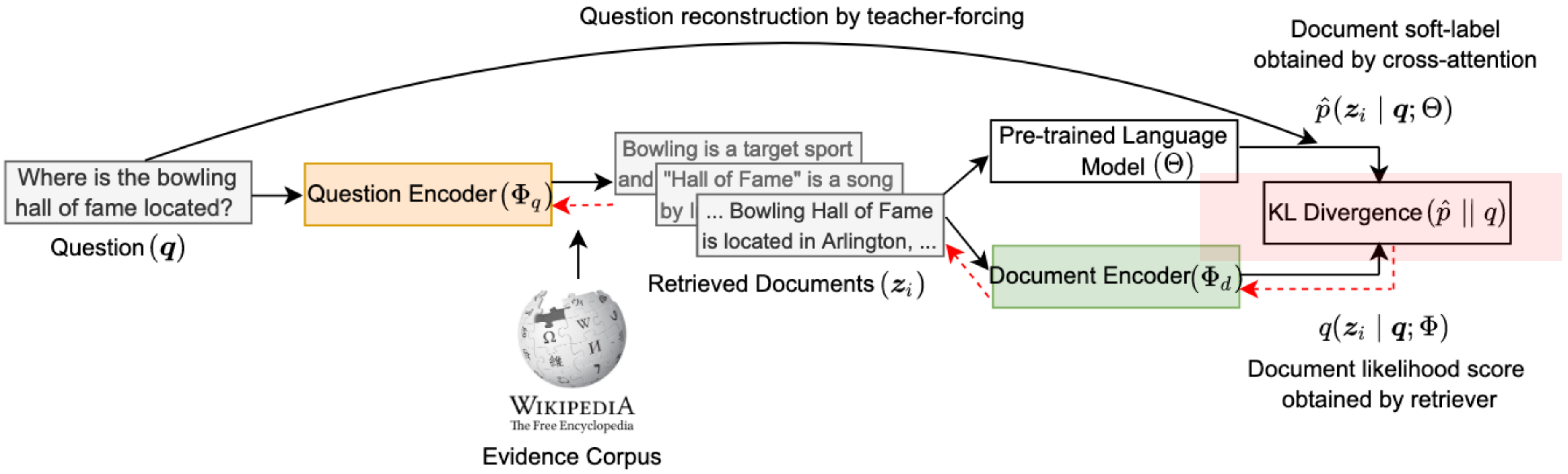
$$\hat{p}(\boldsymbol{z}_i \mid \boldsymbol{q}, \mathcal{Z}) = \frac{\exp(\log p(\boldsymbol{z}_i \mid \boldsymbol{q}; \Theta))}{\sum_{j=1}^{K} \exp\left(\log p(\boldsymbol{z}_j \mid \boldsymbol{q}; \Theta)\right)}$$



ART: Training Details

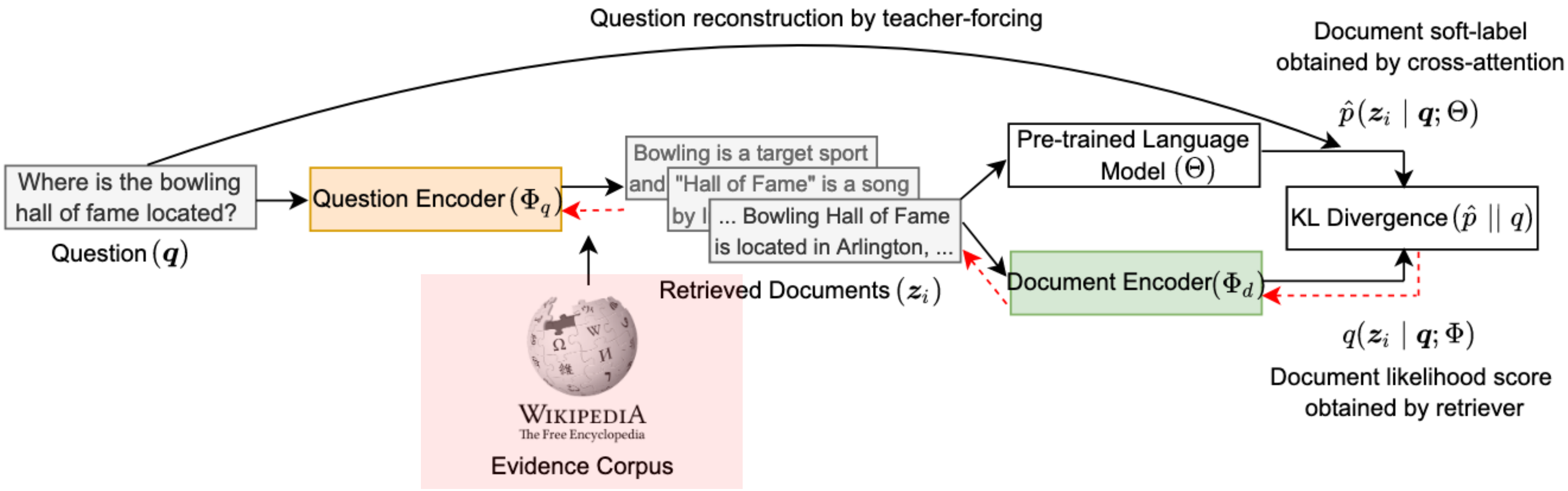
Step 4: Loss calculation and backpropagation

$$\mathcal{L}(\Phi) = \frac{1}{|\mathcal{T}|} \sum_{\boldsymbol{q} \in \mathcal{T}} \mathbb{KL}(\hat{p}(\boldsymbol{z}_i \mid \boldsymbol{q}, \mathcal{Z}) \,||\, q(\boldsymbol{z}_i \mid \boldsymbol{q}, \mathcal{Z}; \Phi))$$



ART: Training Details

Step 5: Periodically update (stale) evidence embeddings



Experiments: Passage Retrieval for QA

We follow DPR-paper experiments style

QA datasets

1. Squad-Open (~78K)
2. TriviaQA (~79K)
3. NQ-Open: Natural Questions (~79K)
4. WebQ: WebQuestions (~3K)

Training Details:

- Relevance scorer **PLM**: T0-3B
- Top-32 documents retrieval
- Batch Size: 64
- GPUs: 8 / 16 A100

Evidence:

- English Wikipedia (2018)
- Each article is segmented into 100 words documents (or passages)
- ~21M documents

Baselines

- **Unsupervised**: trained using Wikipedia text or non-trainable
  1. BM25
  2. Contriever

- **Supervised:** trained using aligned question-document pairs
  1. DPR
  2. ANCE

Contriever: Unsupervised Dense Information Retrieval with Contrastive Learning, Izacard et al., 2022.
ANCE: Approximate Nearest Neighbor Negative Contrastive Learning for Dense Passage Retrieval, Xiong et al., 2021

Results: QA Passage Retrieval

Evaluation Metric

Top-20 accuracy: fraction of questions for which the answer span exists in one of the top-20 documents

Retriever	Zero-Shot	SQuAD-Open	TriviaQA	NQ-Open	WebQ
BM25	✓	71.1	76.4	62.9	62.4
Contriever	✓	63.4	73.9	67.9	65.7
DPR		63.2	79.4	78.4	73.2
ANCE		-	80.3	**81.9**	-
ART	✓	**75.3**	**82.9**	**81.6**	**75.7**

- ART outperforms previous unsupervised methods by **4-12 points**
- ART matches or exceeds performance of supervised models

Results: Single Retriever Training

Evaluation Metric: top-20 accuracy

Retriever	Training Dataset	SQuAD-Open	TriviaQA	Web Questions
BM25	-	71.1	76.4	62.4
		Training on One Dataset		
DPR	NQ-Open	48.9	69.0	68.8
ART	NQ-Open	68.4	79.8	73.4
		Training on All Datasets		
DPR	Multi	51.6	78.8	75.0
ART	Multi	**74.7**	**82.9**	**76.6**

- Improved transfer results on datasets not trained on
- Single model trained on all datasets achieves good results

Real-User Questions

NQ-Open: All questions contain short answers

NQ-Full: Real user questions; **practical setting**

- Short answers
- Long answers (e.g., paragraphs)
- Yes / No answers
- **Questions do not have the answer in Wikipedia**
- **Ambiguous questions**

	# Questions
NQ-Open	~79K
NQ-Full	~307K

In NQ-Full, > 51% of questions are unanswerable

Natural Questions: a Benchmark for Question Answering Research, Kwiatkowski et al., TACL 2019.

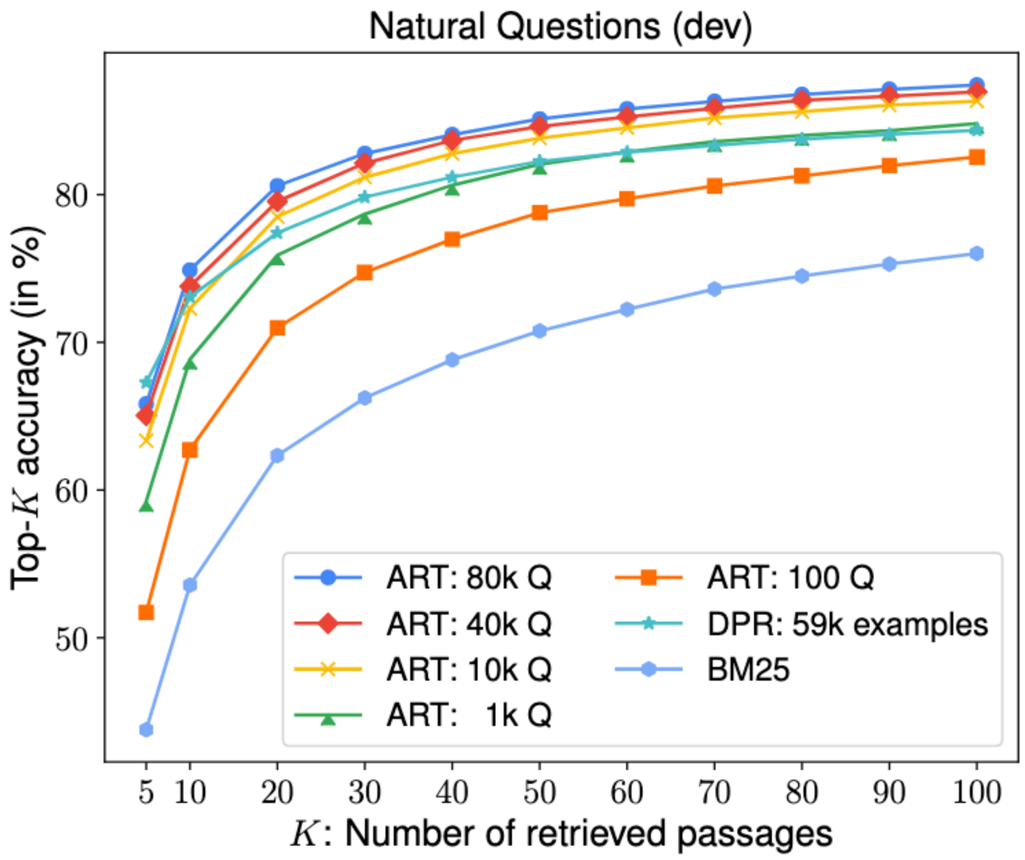
Results: Robustness to Unanswerable Questions

Evaluation Metric: Top-20 accuracy

Method	Training Dataset	SQuAD-Open	TriviaQA	WebQ
Training on answerable questions				
BM25	-	**71.1**	76.4	62.4
DPR	NQ-Open	48.9	69.0	68.8
ART	NQ-Open	68.4	79.8	73.4
Training on a mix of answerable and unanswerable questions				
ART	NQ-Full	69.4	**80.3**	**74.3**

- ART can be trained on both answerable and unanswerable questions
- Small gain in performance

Results: Sample Efficiency



- ART is more sample efficient than DPR
- Outperforms BM25 with just 100 questions

Analysis: Impact of PLM Training

PLM training style

- *Denoising spans*
  - Ex: T5, BART
- *Autoregressive generative training*
  - Ex: GPT, T5-lm-adapt
- *Instruction-tuning*
  - Ex: T0, FLANN

Language Model (Θ)	**NQ-Open (dev)**			
	Top-1	Top-5	Top-20	Top-100
Models trained using Denoising Masked Spans				
T5-base (250M)	12.8	30.9	47.8	63.0
T5-xl (3B)	25.0	53.9	74.4	85.3
Models trained using Language Modeling Objective				
T5-lm-adapt (250M)	29.4	56.6	74.4	84.7
T5-lm-adapt (800M)	30.9	59.1	76.5	85.9
T5-lm-adapt (3B)	31.8	61.0	77.9	86.5
Model trained using Natural Language Instructions				
T0-3B	**36.7**	**65.8**	**80.6**	**87.4**

- **Instruction-tuned language models are the most effective as relevance scorers**
- Accuracy improves with larger PLMs

Analysis: Why Top-K Document Retrieval

In the top-K documents, we include:

- **U**: uniformly sampled document
- **P**: Positive document
- **N**: hard-negative document
- **IB**: In-batch training

P	N	U	IB	Top-1	Top-5	Top-20	Top-100
0	0	32	✗	6.0	16.6	30.8	46.7
1	0	31	✗	31.8	58.9	74.8	84.4
1	1	30	✗	33.7	61.0	76.0	85.5
1	1	0	✓	32.6	59.5	75.1	84.9
Top-32 passages				**36.7**	**65.8**	**80.6**	**87.4**

Retrieving top-K documents during training is **crucial** for good performance

Analysis: Limitations

A Closer Comparison with Supervised Models

Supervised Baselines

- **DPR**: finetunes dual-encoder
- **EMDR^2**: finetunes both PLM and dual-encoder using end-to-end training

Retriever	Top-1	Top-5	Top-20	Top-100
NQ-Open (dev)				
DPR	**50.1**	**69.6**	79.1	85.5
EMDR^2	**55.3**	**74.9**	83.1	88.0
ART	37.6	66.8	81.0	87.8
TriviaQA (dev)				
DPR	59.6	74.4	81.1	85.9
EMDR^2	**63.7**	78.0	83.7	87.4
ART	58.3	77.5	83.7	87.5

ART lags in top-1 and top-5 retrieval accuracy

EMDR^2: End-to-End Training of Multi-Document Reader and Retriever for Open-Domain Question Answering, Sachan et al., NeurIPS 2021

Analysis: Importance of Question Reconstruction

Option 1: Likelihood of autoregressive question reconstruction

$$\log p(\boldsymbol{z}_i \mid \boldsymbol{q}; \Theta) \propto \frac{1}{|\boldsymbol{q}|} \sum_t \log p(q_t \mid \boldsymbol{q}_{<t}, \boldsymbol{z}_i; \Theta)$$

Option 2: Likelihood of autoregressive document reconstruction

$$\log p(\boldsymbol{z}_i \mid \boldsymbol{q}; \Theta) \propto \frac{1}{|\boldsymbol{z_i}|} \sum_t \log p(z_t \mid \boldsymbol{z}_{i<t}, \boldsymbol{q}; \Theta)$$

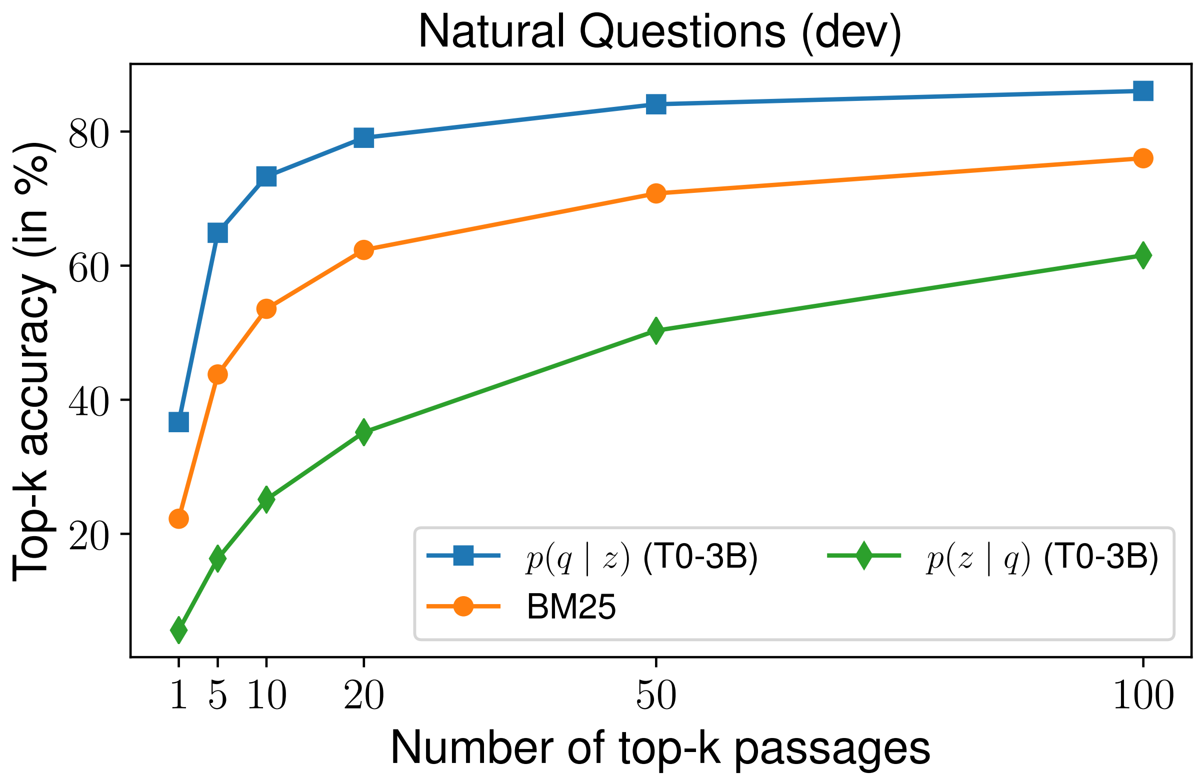
Analysis: Importance of Question Reconstruction

Experiment: re-rank top-1000 documents from BM25 using T0-3B

Baseline: BM25

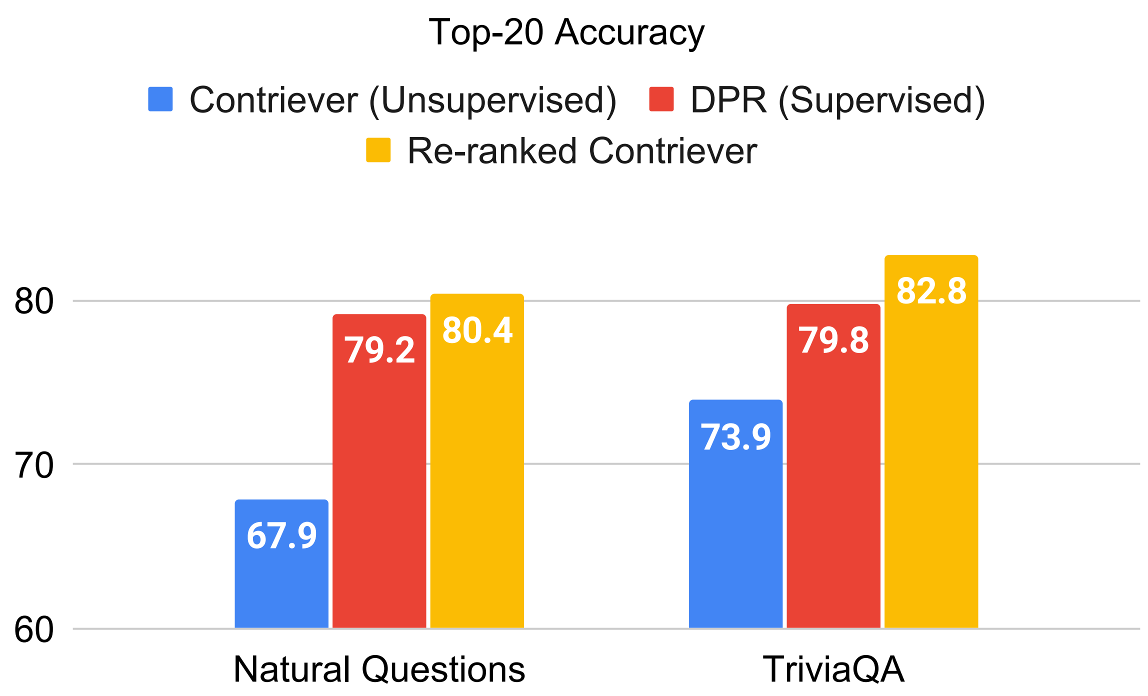
p(q | z) : Likelihood of question tokens

p(z | q) : Likelihood of document tokens



Question reconstruction i.e., **p(q | z)** > BM25 > Document reconstruction i.e., **p(z | q)**

Unsupervised Re-ranking using Question Reconstruction



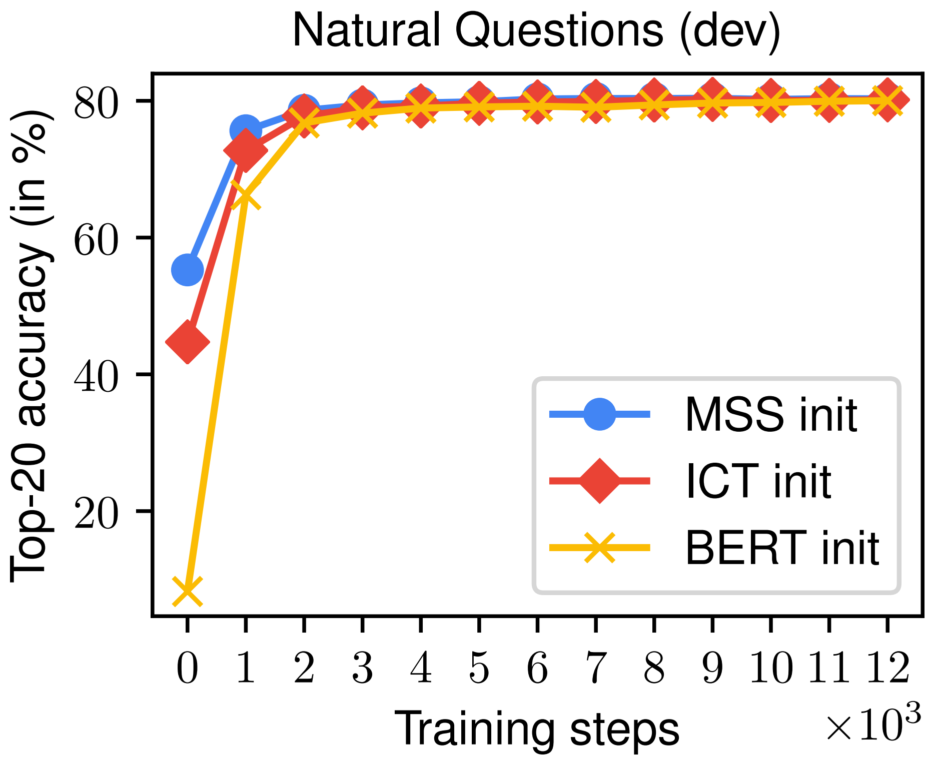
- Re-ranking top-1000 passages from Contriever using T0-3B PLM
- Better performance than DPR

Improving Passage Retrieval with Zero-Shot Question Generation, Sachan et al., EMNLP 2022

Analysis: Effect of Retriever Initialization

Retriever is initialized using

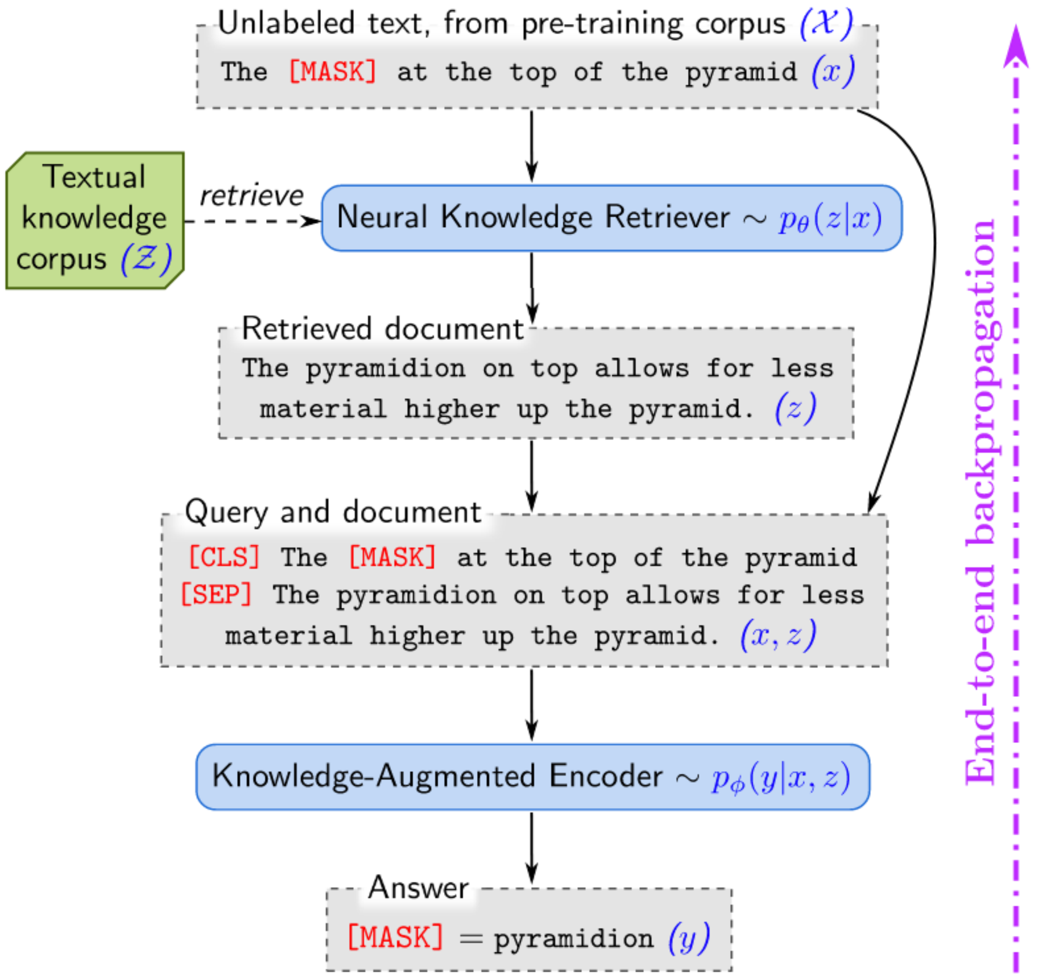
1. BERT
2. Inverse Cloze Task (ICT)
3. Masked Salient Spans (MSS)



ART is robust to retriever initialization

ICT: Latent Retrieval for Weakly Supervised Open Domain Question Answering, Lee et al., ACL 2019.
MSS: End-to-End Training of Neural Retrievers for Open-Domain Question Answering, Sachan et al., ACL 2021

Review: Pre-training using Masked Salient Span (MSS)



Masked Salient Span (MSS) task:

- Identify salient spans such as named entities in sentences
- Mask salient spans and predict using retrieved documents

REALM: Retrieval-Augmented Language Model Pre-Training, Guu et al., ICML 2020

Discussion: MSS or ART style training ?

Top-20 accuracy

Model	Training Task	NQ-Open	TriviaQA
REALM	MSS	59.8	68.2
ART	Question Reconstruction	81.6	82.9

- MSS is **not an ideal proxy task to train retriever**
- ART based on questions is a **promising** alternative

REALM: Retrieval-Augmented Language Model Pre-Training, Guu et al., ICML 2020

Some Ideas for Future Work

1. Application to low-resource settings such as *multi-lingual and cross-lingual retrieval*

2. Application to cross-modality retrievers such as *image and code retrieval using text*

3. New approaches to improve **top-1** and **top-5** retrieval accuracy

Questions / Discussion

Paper: https://arxiv.org/abs/2206.10658

Code: https://github.com/DevSinghSachan/art

E-mail: sachan.devendra@gmail.com

Thank you!