End-to-End Training of Multi-Document Reader and Retriever for Open-Domain QA

- Devendra Singh Sachan

PhD student @ Mila and McGill University

Devendra Sachan

Siva Reddy

Will Hamilton

Chris Dyer

Dani Yogatama

Problem Setup: Open-Domain QA

- **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

Five Australian prime ministers who governed during World War 2 → Arthur Fadden, John Curtin, Frank Forde, Robert Menzies, Ben Chifley

WIKIPEDIA
The Free Encyclopedia

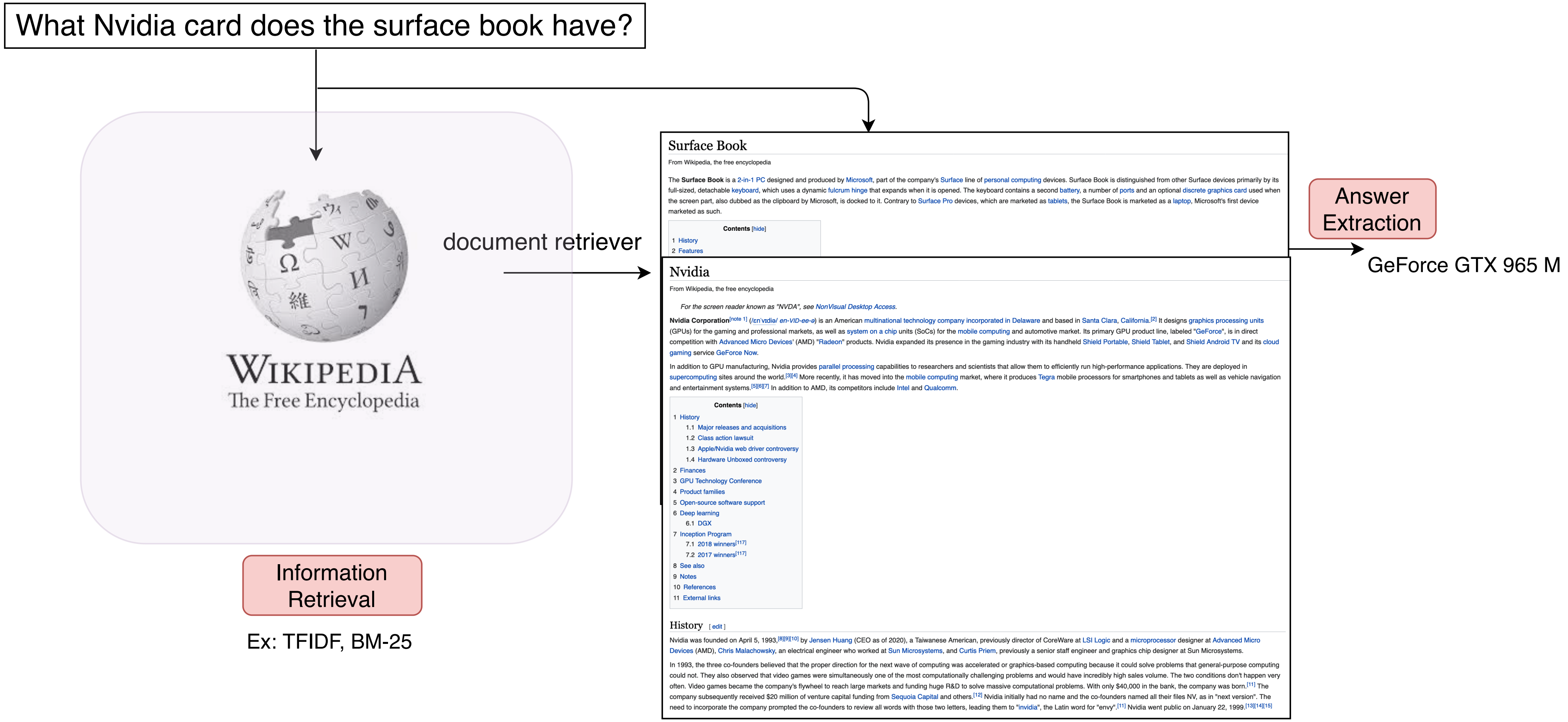
Chen et al., 2017. Reading Wikipedia to Answer Open-domain Questions

Background: Open-Domain QA

Two-stage approach

Stage 1. Document retriever from evidence

Stage 2. Answer extraction



Background: Neural Models for Open-Domain QA

Stage 1: Trainable Information Retrieval

Stage 2: Trainable Answer Extraction

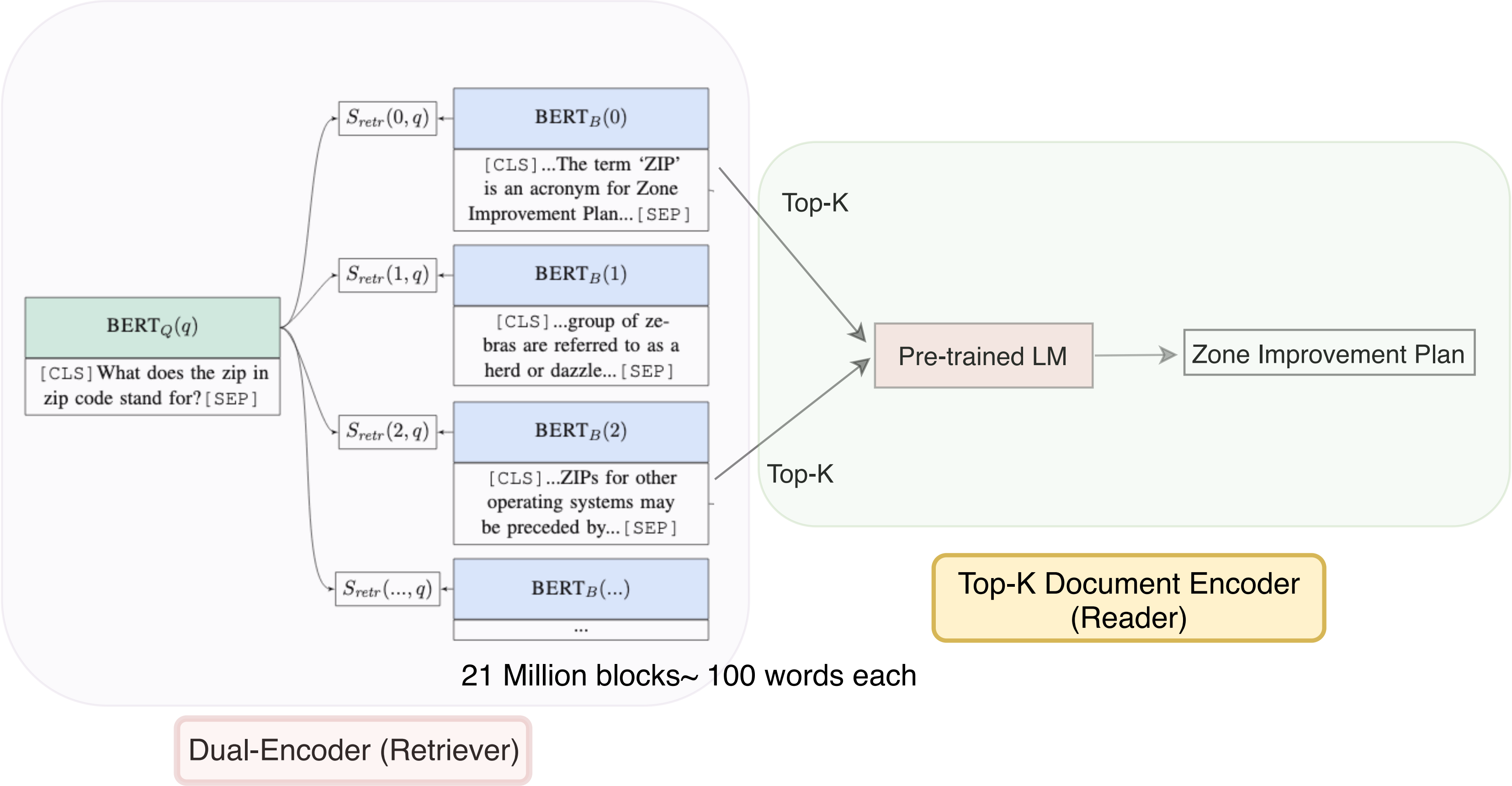


Image from "Lee et al., 2019. Latent Retrieval for Weakly Supervised Open Domain Question Answering"

EMDR^2: End-to-End Training of Multi-Document Reader and Retriever

Modeling Components

- **Retriever**: Dual Encoder
- **Reader / Answer Extractor**: Fusion-in-Decoder (FiD)

EMDR^2: End-to-End Training of Multi-Document Reader and Retriever

Modeling Components

- **Retriever**: Dual Encoder
- **Reader / Answer Extractor**: Fusion-in-Decoder (FiD)

Dual Encoder Retriever

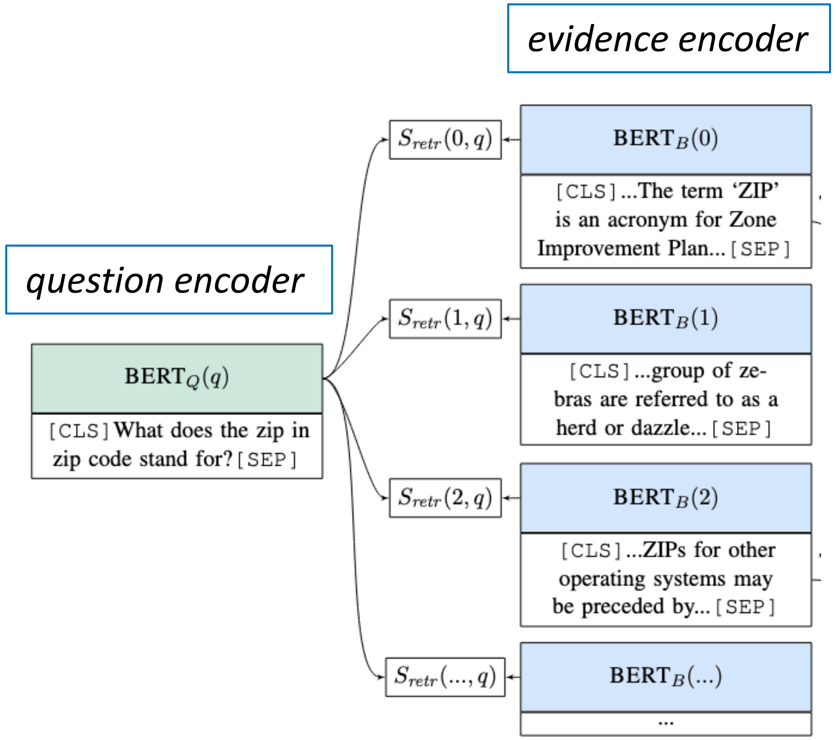
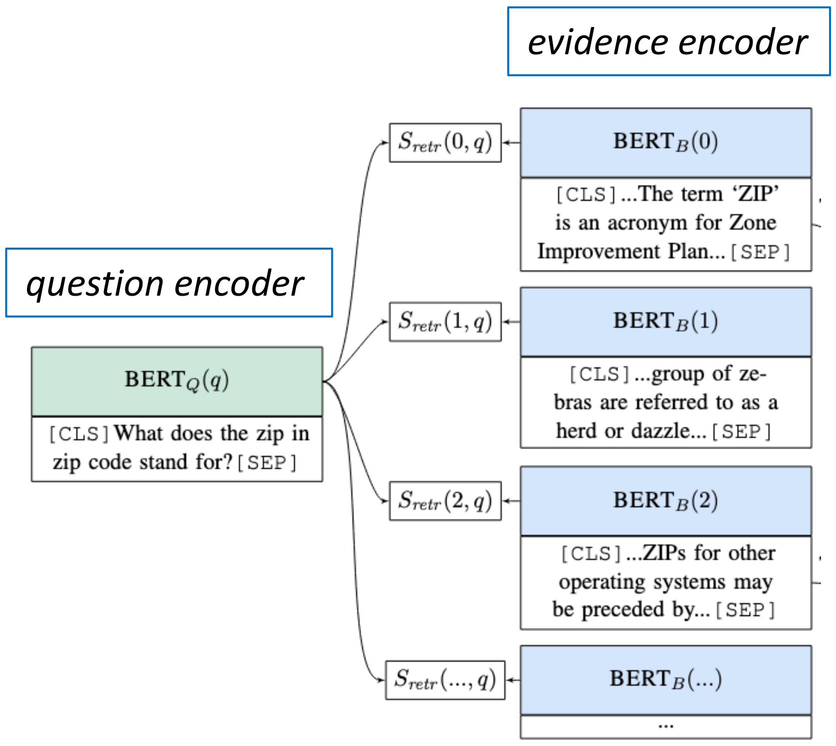


Image from "Lee et al., 2019. Latent Retrieval for Weakly Supervised Open Domain Question Answering"

Dual Encoder Retriever



Evidence Documents

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

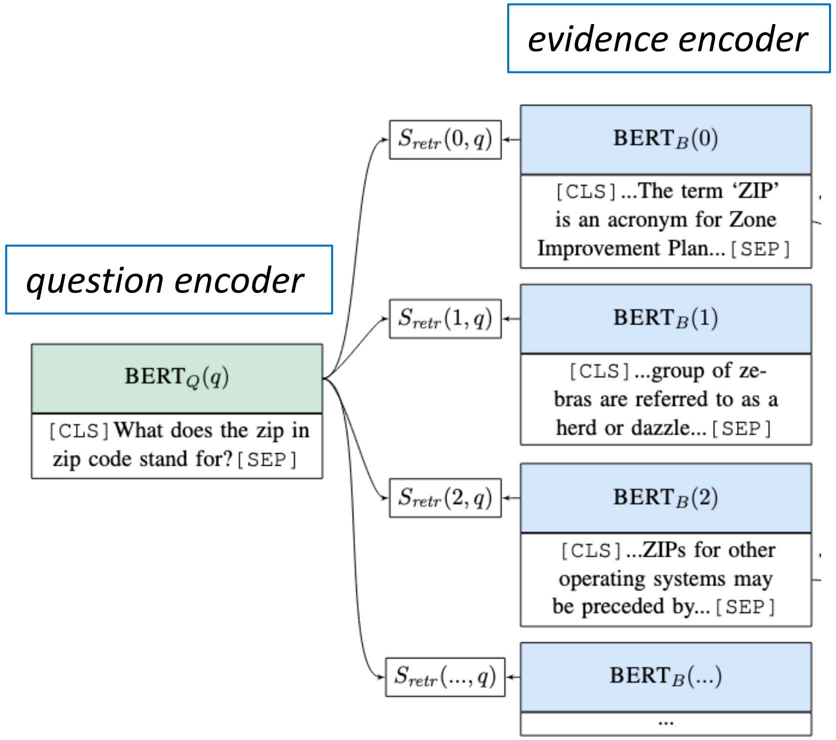
$$\text{score}(\boldsymbol{q}, \boldsymbol{d}_i; \Phi) = f_q(\boldsymbol{q}; \Phi_q)^\top f_d(\boldsymbol{d}_i; \Phi_d)$$

Select Top-K Documents with highest scores

$$\mathcal{Z} = \{\boldsymbol{z}_1, \ldots, \boldsymbol{z}_K\}$$

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

Dual Encoder Retriever



Evidence Documents

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

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

Select Top-K Documents with highest scores

$$\mathcal{Z} = \{\boldsymbol{z}_1, \ldots, \boldsymbol{z}_K\}$$

Dual encoder is initialized with **Inverse Cloze Task (ICT)**

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

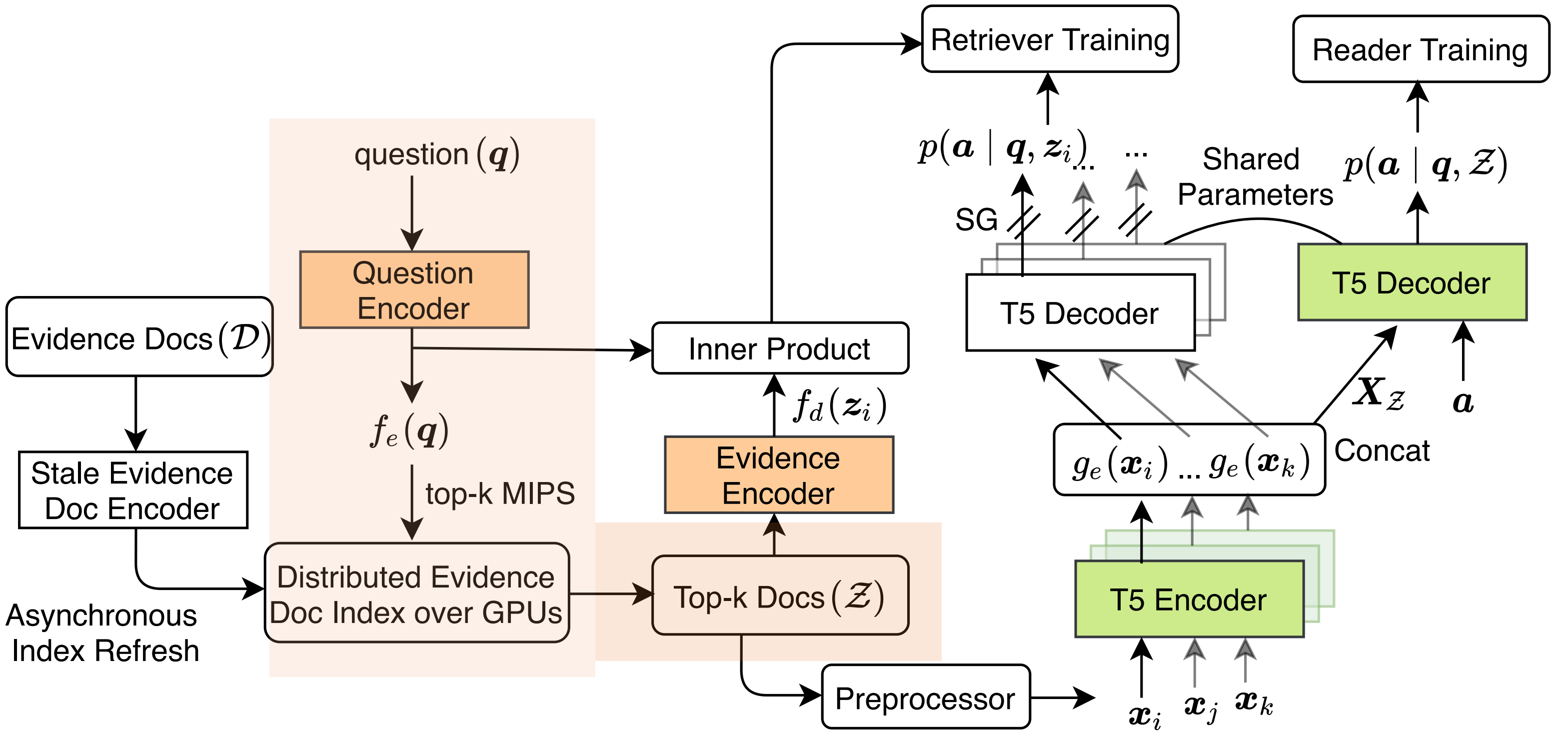
Evidence Documents

- **Evidence**: document collection containing world knowledge.
- We use **English Wikipedia** from Dec 2018 as evidence.
- Split articles into **100 words** long sequences.
  - Shorter sequences -> higher retrieval accuracy
- Overall size = 21 Million sequences

Top-K Documents Retrieval

- **Pre-compute** evidence embeddings with context encoder.
- Distributed evidence storage over 16 GPUs.
  - 1.3 M document embeddings stored in each GPU
- We perform **online retrieval** at every step.
- Retrieval by asynchronous matrix multiplication in multiple GPUs.

EMDR^2: Top-K Retrieval



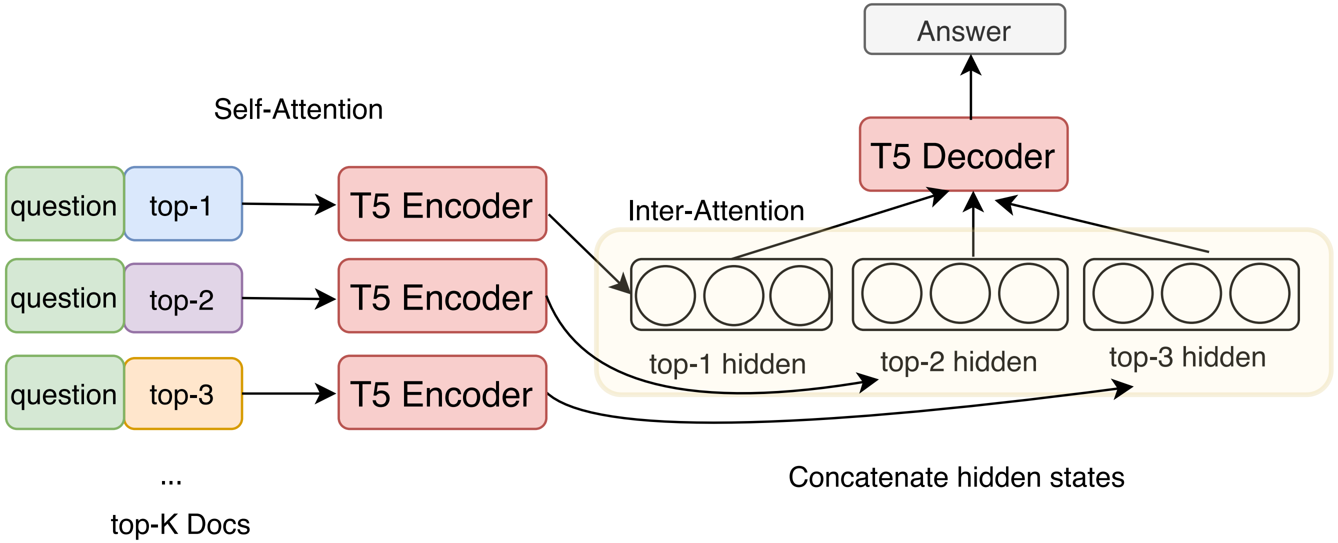
EMDR^2: End-to-End Training of Multi-Document Reader and Retriever

Modeling Components

- **Retriever**: Dual Encoder
- **Reader / Answer Extractor**: **Fusion-in-Decoder (FiD)**

Multi-Document Reader: Fusion-in-Decoder

FiD: generative approach for answer extraction based on T5



Encoder Input

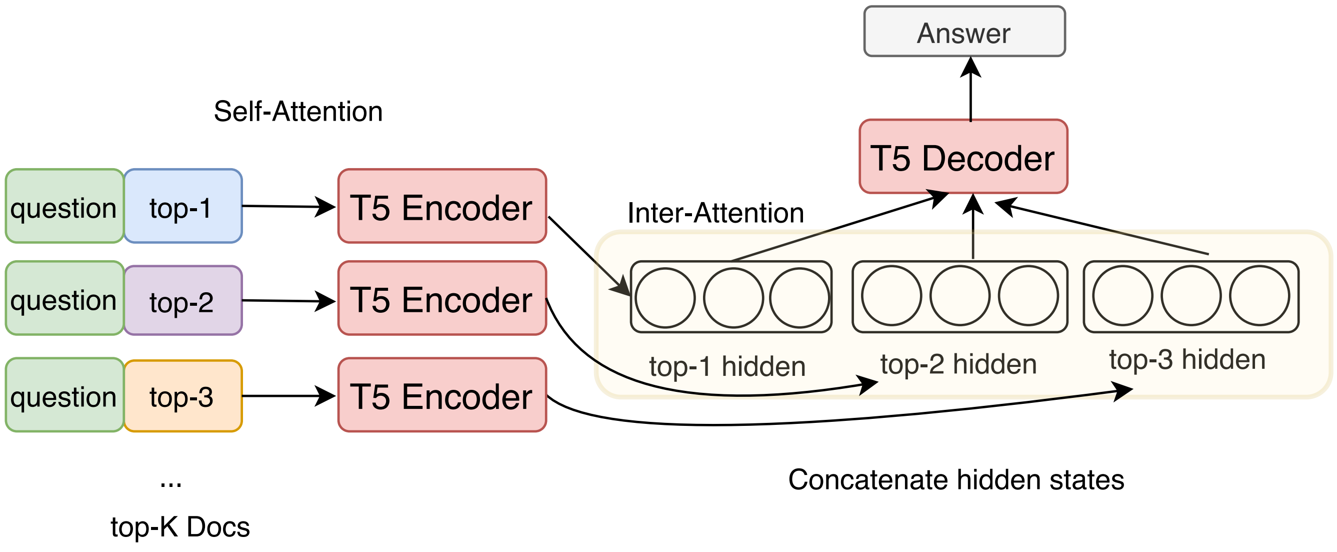
for each Top-K doc:

$$x_k = \text{[CLS]}\, q\, \text{[SEP]}\, t_{z_k}\, \text{[SEP]}\, z_k\, \text{[SEP]}$$

q = question
z_k = top-K document
t_z = title of top-K document

"Izacard et al., 2021. Leveraging Passage Retrieval with Generative Models for Open Domain Question Answering"

Fusion-in-Decoder: Self-Attention

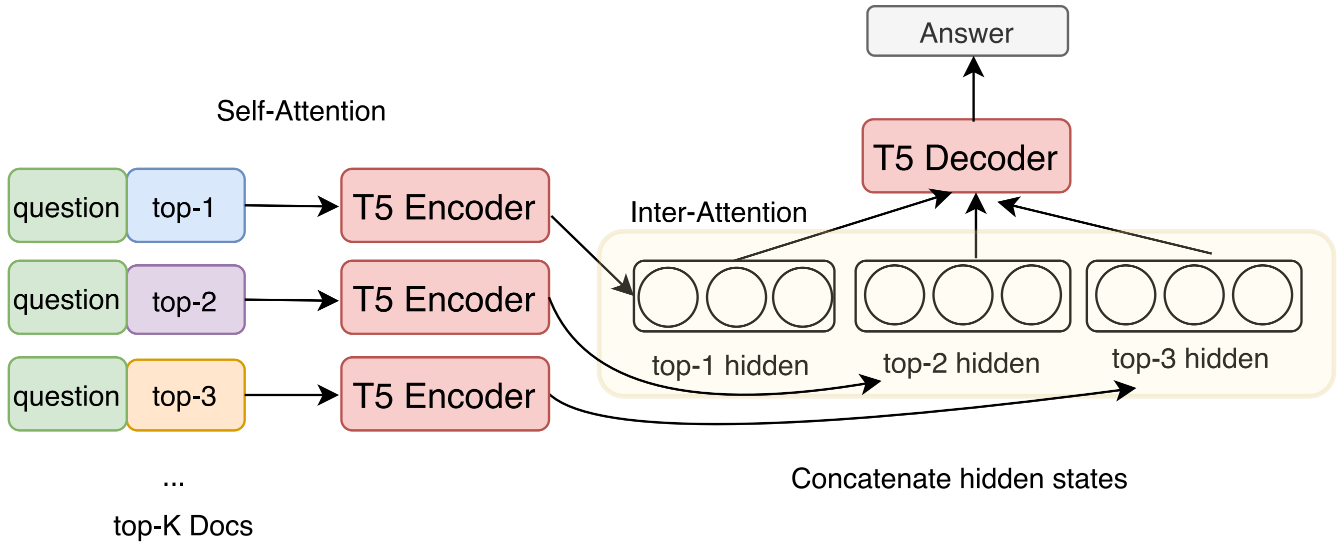


$$\boldsymbol{x}_k = [\text{CLS}]\,\boldsymbol{q}\,[\text{SEP}]\,\boldsymbol{t}_{\boldsymbol{z}_k}\,[\text{SEP}]\,\boldsymbol{z}_k\,[\text{SEP}]$$

Independent self-attention over each x_k

"Izacard et al., 2021. Leveraging Passage Retrieval with Generative Models for Open Domain Question Answering"

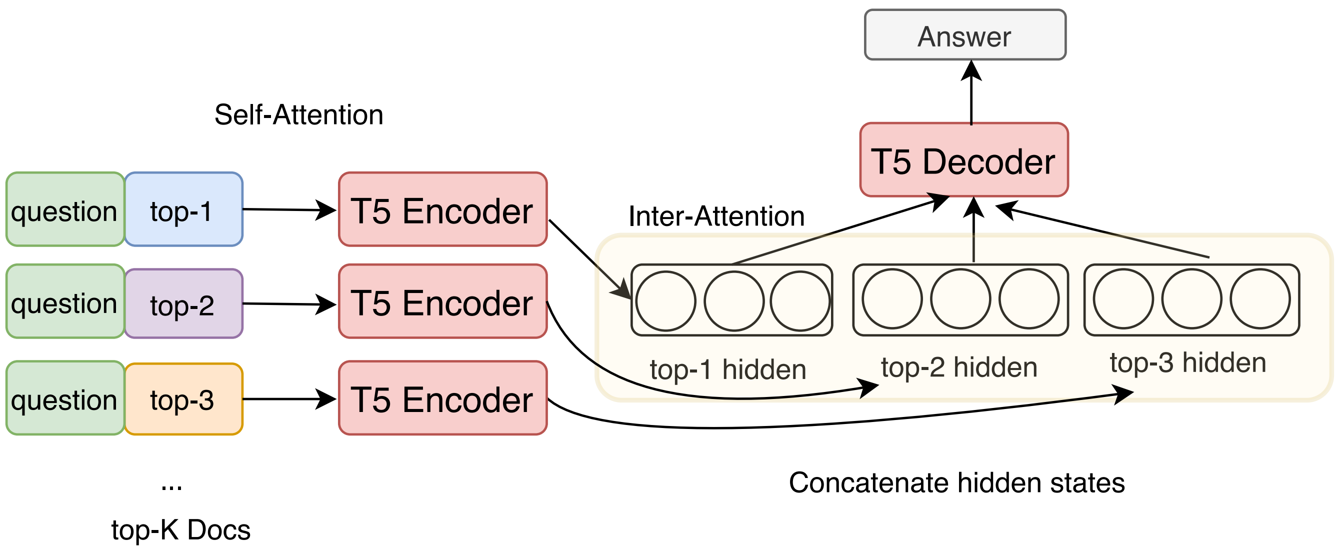
Fusion-in-Decoder: Inter-Attention



Concatenate the encoder representations for decoder's inter-attention

$$\mathbf{X}_{\mathcal{Z}} = [g_e(\boldsymbol{x}_1); \ldots ; g_e(\boldsymbol{x}_K)] \in \mathbb{R}^{(N \times K) \times H}$$

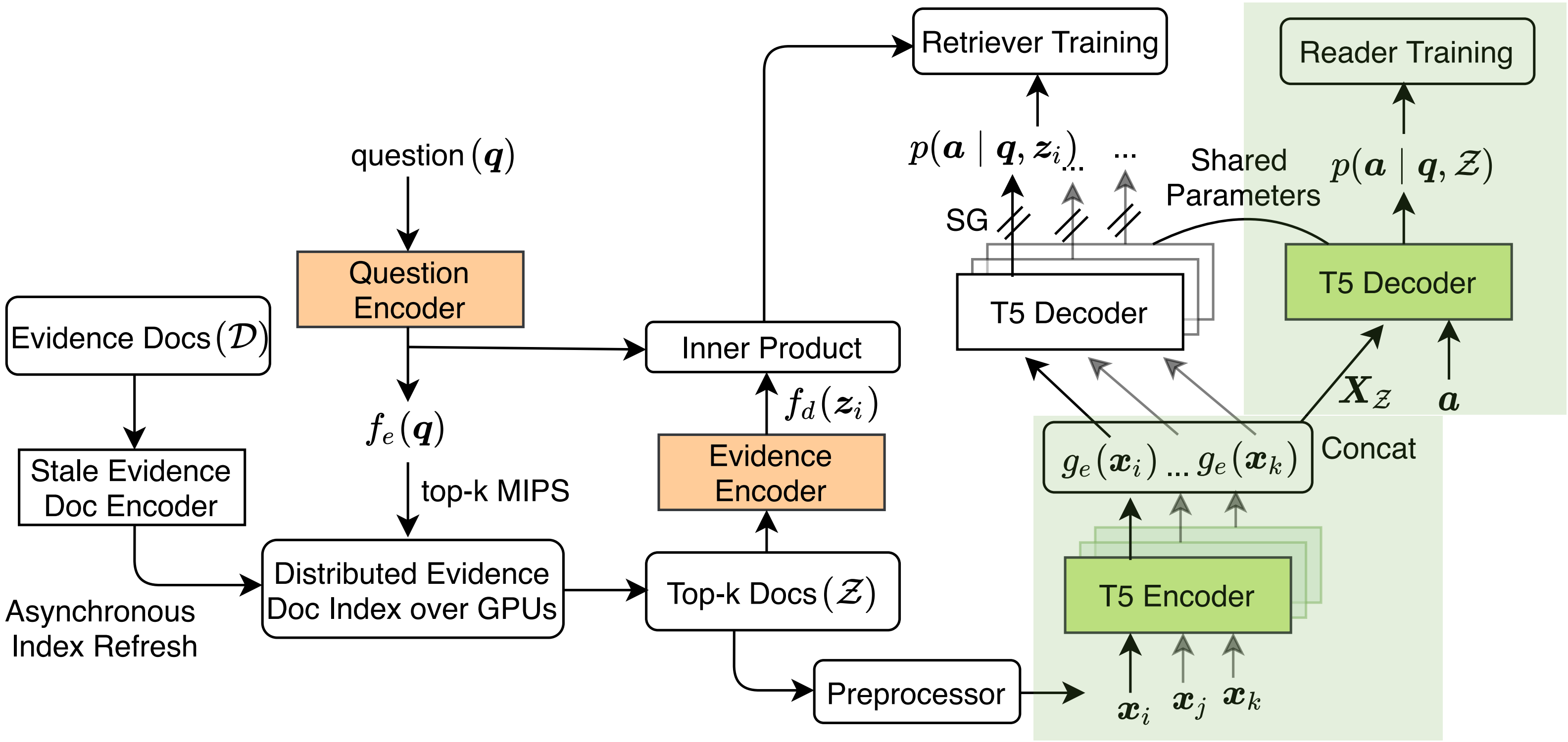
Fusion-in-Decoder: Training



Autoregressively train the model

$$p(\boldsymbol{a} \mid \boldsymbol{q}, \mathcal{Z}; \Theta) = \prod_{t=1}^{T} p\left(a_t \mid \boldsymbol{a}_{<t}, \boldsymbol{q}, \mathcal{Z}; \Theta\right)$$

EMDR^2: Fusion-in-Decoder



End-to-End Training: Formulation

Marginal Likelihood

$$p(\boldsymbol{a} \mid \boldsymbol{q}; \Theta, \Phi) = \sum_{\mathcal{Z} \in \mathscr{S}} \underbrace{p(\boldsymbol{a} \mid \boldsymbol{q}, \mathcal{Z}; \Theta)}_{\text{FiD}} \underbrace{p(\mathcal{Z} \mid \boldsymbol{q}; \Phi)}_{\text{Dual Encoder}}$$

- The **set** of retrieved documents $\mathcal{Z}$ is a latent variable.

End-to-End Training: Formulation

Marginal Likelihood

$$p(\boldsymbol{a} \mid \boldsymbol{q}; \Theta, \Phi) = \sum_{\mathcal{Z} \in \mathscr{S}} \underbrace{p(\boldsymbol{a} \mid \boldsymbol{q}, \mathcal{Z}; \Theta)}_{\text{FiD}} \underbrace{p(\mathcal{Z} \mid \boldsymbol{q}; \Phi)}_{\text{Dual Encoder}}$$

- The **set** of retrieved documents $\mathcal{Z}$ is a latent variable.
- All possible values of $\mathcal{Z}$ are **combinatorial** in nature.

$$\mathscr{S} = \binom{M}{K}$$

End-to-End Training: Formulation

For **one** particular value of $\mathcal{Z}$, log-likelihood becomes

$$\begin{aligned}\log p(\boldsymbol{a} \mid \boldsymbol{q}; \Theta) &\approx \log p(\boldsymbol{a} \mid \boldsymbol{q}, \mathcal{Z}; \Theta) p(\mathcal{Z} \mid \boldsymbol{q}; \Phi) \\ &\approx \log p(\boldsymbol{a} \mid \boldsymbol{q}, \mathcal{Z}; \Theta) + \log p(\mathcal{Z} \mid \boldsymbol{q}; \Phi)\end{aligned}$$

End-to-End Training: Formulation

Log-Likelihood

$$\mathcal{L} = \log \underbrace{p(\boldsymbol{a} \mid \boldsymbol{q}, \mathcal{Z}_{\text{reader}}; \Theta)}_{\text{FiD training}} + \log \underbrace{p(\mathcal{Z}_{\text{retriever}} \mid \boldsymbol{q}; \Phi)}_{\text{Dual Encoder training}}$$

EMDR^2: FiD Training

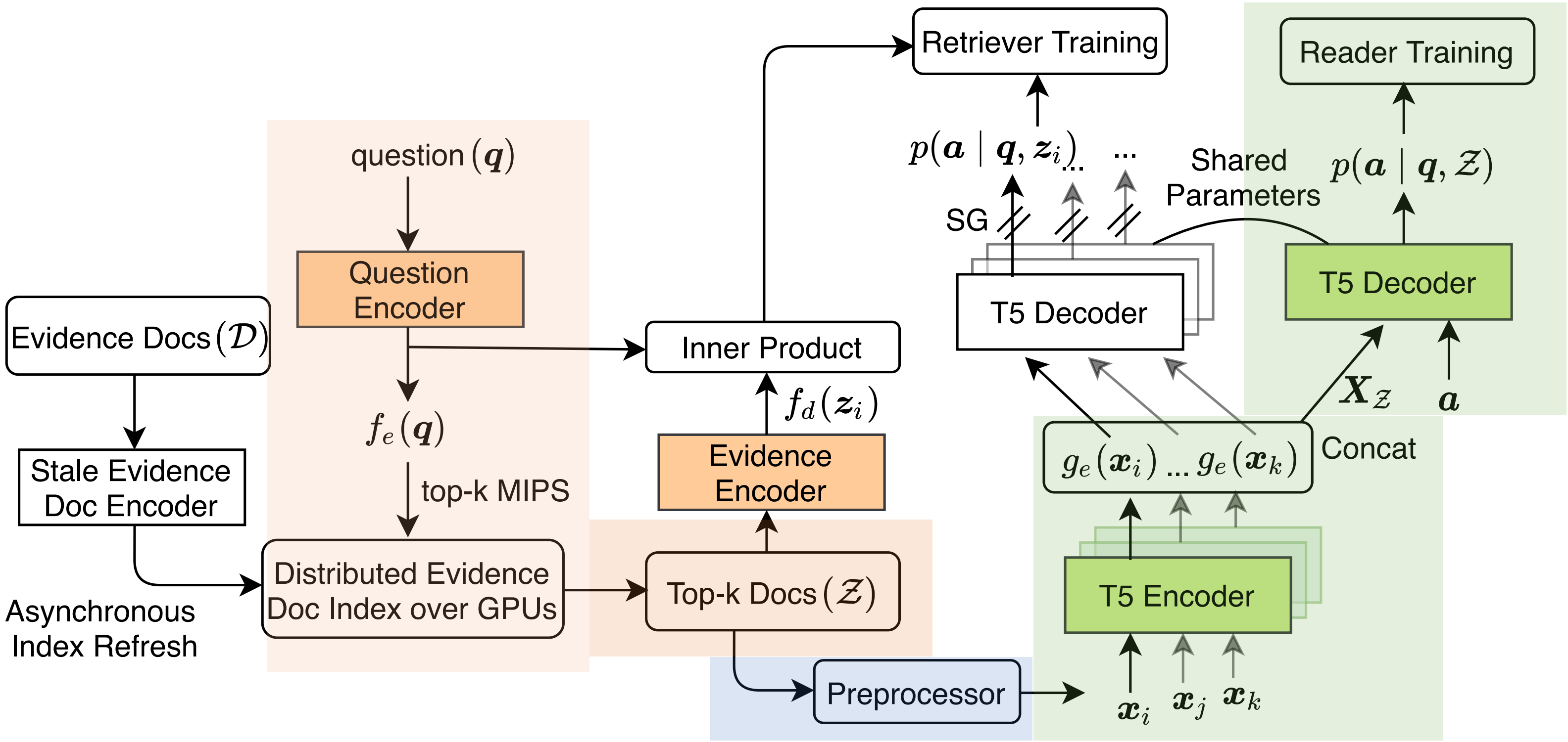
$$\mathcal{L} = \underbrace{\log p(\boldsymbol{a} \mid \boldsymbol{q}, \mathcal{Z}_{\text{reader}}; \Theta)}_{\text{FiD training}} + \log \underbrace{p(\mathcal{Z}_{\text{retriever}} \mid \boldsymbol{q}; \Phi)}_{\text{Dual Encoder training}}$$

- Obtain top-K documents of $\mathcal{Z}$ based on Maximum Inner Product Search (MIPS)

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

- Teacher-forcing training of FiD

EMDR^2: FiD Training



EMDR^2: Retriever Training

$$\mathcal{L} = \underbrace{\log p(\boldsymbol{a} \mid \boldsymbol{q}, \mathcal{Z}_{\text{reader}}; \Theta)}_{\text{FiD training}} + \underbrace{\log p(\mathcal{Z}_{\text{retriever}} \mid \boldsymbol{q}; \Phi)}_{\text{Dual Encoder training}}$$

- Just optimizing the second term leads to poor results
- We need some form of supervision for retriever training, which comes from the answer (a)

EMDR^2: Retriever Training

We use posterior as it contains dependence on answer (a)

Simplifying Assumption: max probability of a set = max of the total probability of its elements

$$\max p(\mathcal{Z}_{\text{retriever}} \mid \boldsymbol{q}, \boldsymbol{a}; \Theta, \Phi) = \max \sum_{k=1}^{K} p(\boldsymbol{z}_k \mid \boldsymbol{q}, \boldsymbol{a}; \Theta, \Phi)$$

EMDR^2: Retriever Training

We use posterior for better training as it contains dependence on answer (a)

From **Conditional Bayes Rule:**

$$p(\boldsymbol{z}_k \mid \boldsymbol{q}, \boldsymbol{a}; \Theta, \Phi) \propto \underbrace{p(\boldsymbol{a} \mid \boldsymbol{q}, \boldsymbol{z}_k; \Theta)}_{\text{T5}} \underbrace{p(\boldsymbol{z}_k \mid \boldsymbol{q}; \Phi)}_{\text{Dual Encoder}}$$

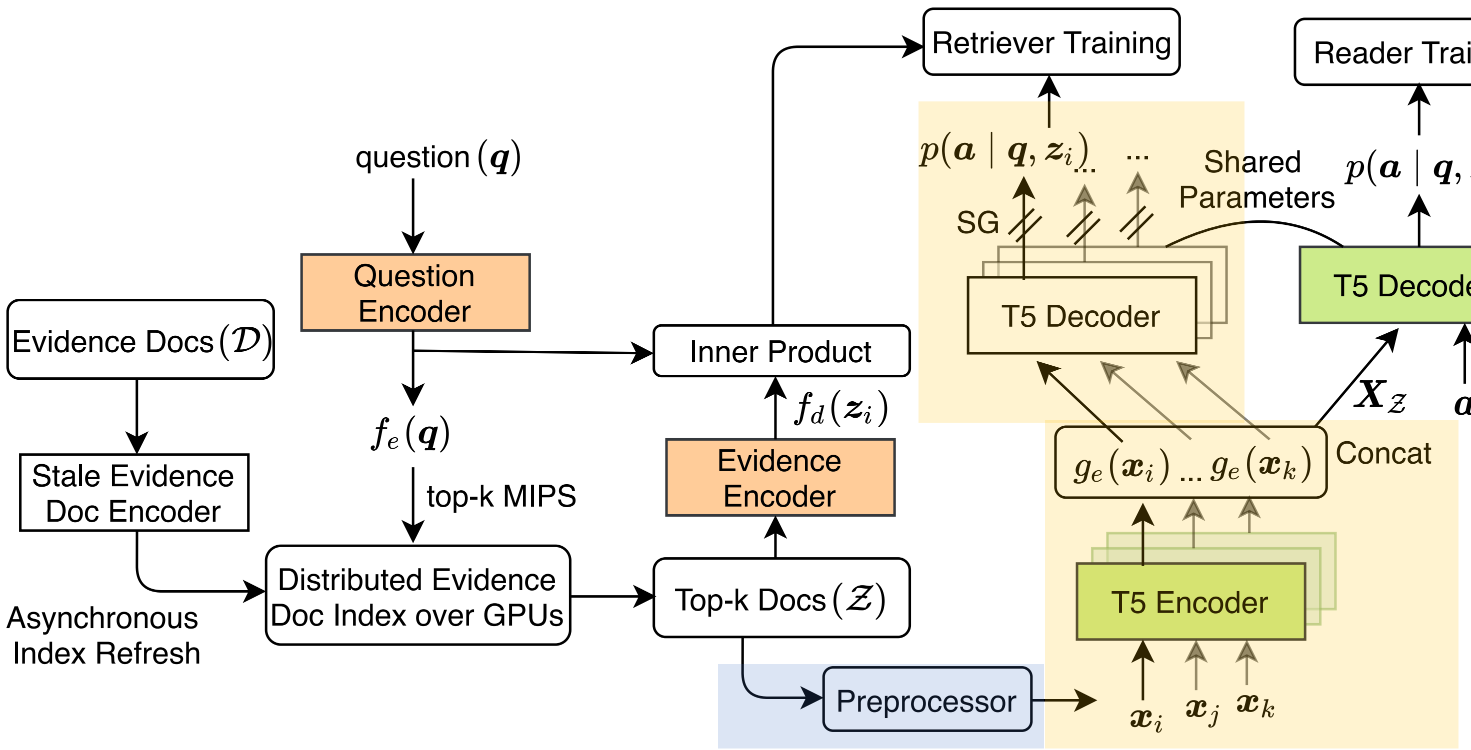
EMDR^2: Retriever Training

From **Conditional Bayes Rule:**

$$p(\boldsymbol{z}_k \mid \boldsymbol{q}, \boldsymbol{a}; \Theta, \Phi) \propto \underbrace{p(\boldsymbol{a} \mid \boldsymbol{q}, \boldsymbol{z}_k; \Theta)}_{\text{T5}} \underbrace{p(\boldsymbol{z}_k \mid \boldsymbol{q}; \Phi)}_{\text{Dual Encoder}}$$

First part can be computed from T5

EMDR^2: Retriever Training



EMDR^2: Retriever Training

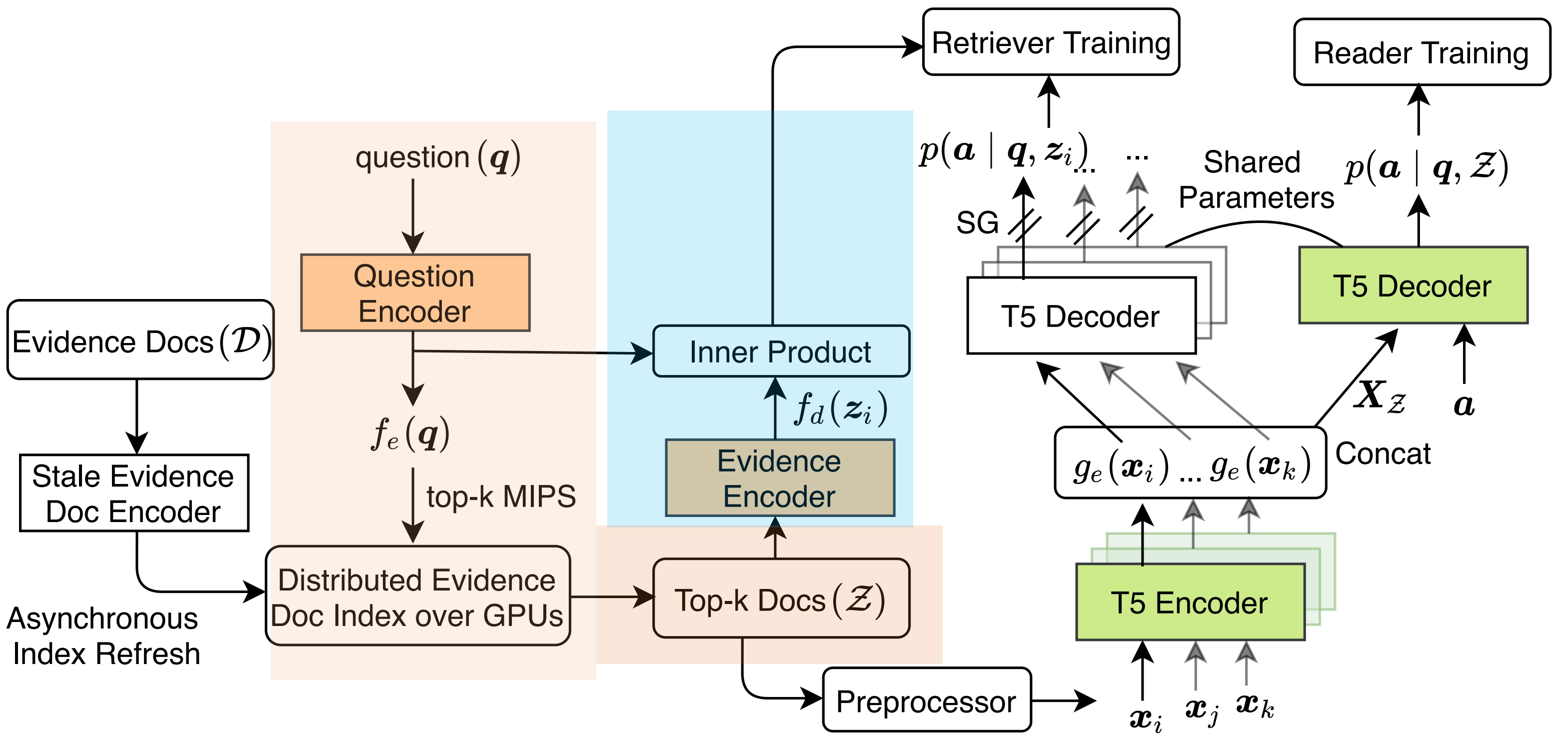
Probability of a document z_k from dual encoder

$$p(\boldsymbol{z}_k \mid \boldsymbol{q}, \boldsymbol{a}; \Theta, \Phi) \propto \underbrace{p(\boldsymbol{a} \mid \boldsymbol{q}, \boldsymbol{z}_k; \Theta)}_{\text{T5}} \underbrace{p(\boldsymbol{z}_k \mid \boldsymbol{q}; \Phi)}_{\text{Dual Encoder}}$$

Apply a softmax with temperature over the top-K scores

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

EMDR²: Retriever Training



EMDR^2 Training Objective

$$\mathcal{L} = \underbrace{\log p(\boldsymbol{a} \mid \boldsymbol{q}, \mathcal{Z}_{\text{top-}K}; \Theta)}_{\text{T5 training}} + \underbrace{\log \sum_{k=1}^{K} \mathbb{SG}\left(p(\boldsymbol{a} \mid \boldsymbol{q}, \boldsymbol{z}_k; \Theta)\right) p(\boldsymbol{z}_k \mid \boldsymbol{q}, \mathcal{Z}_{\text{top-}K}; \Phi)}_{\text{Dual Encoder training}}$$

SG: Stop Gradient operation i.e., no backpropagation

EMDR^2 Training Objective: EM View

E Step:

1. Obtain top-K documents $\mathcal{Z}$ based on current retriever parameters

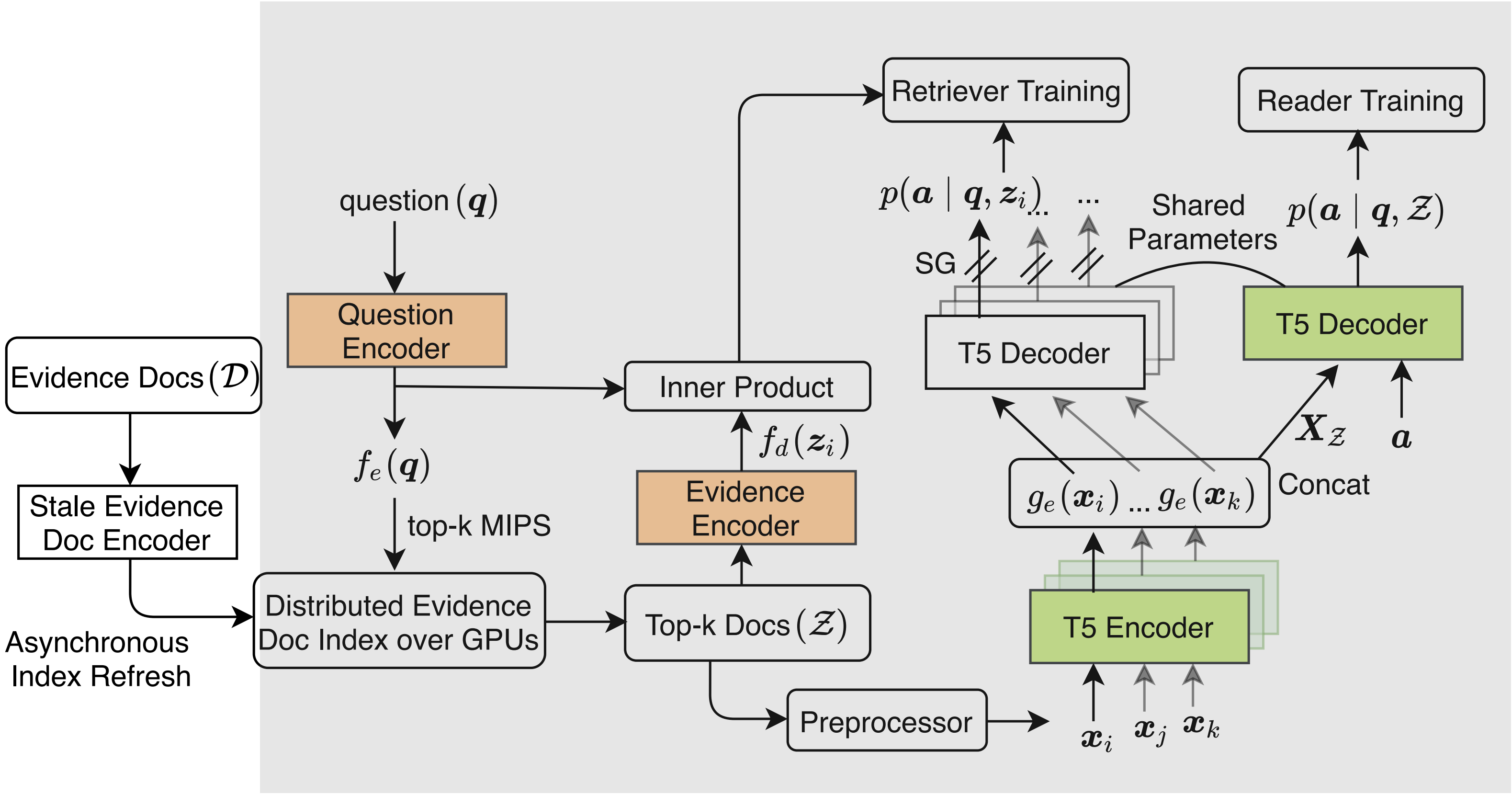
$$\text{score}(\boldsymbol{q}, \boldsymbol{d}_i; \Phi) = f_q(\boldsymbol{q}; \Phi_q)^\top f_d(\boldsymbol{d}_i; \Phi_d)$$

2. Obtain $p(\boldsymbol{a} \mid \boldsymbol{q}, \boldsymbol{z}_k; \Theta)$ based on current T5 parameters

M Step:

$$\mathcal{L} = \underbrace{\log p(\boldsymbol{a} \mid \boldsymbol{q}, \mathcal{Z}_{\text{top-}K}; \Theta)}_{\text{T5 training}} + \underbrace{\log \sum_{k=1}^{K} \mathbb{SG}\left(p(\boldsymbol{a} \mid \boldsymbol{q}, \boldsymbol{z}_k; \Theta)\right) p(\boldsymbol{z}_k \mid \boldsymbol{q}, \mathcal{Z}_{\text{top-}K}; \Phi)}_{\text{Dual Encoder training}}$$

EMDR^2: Modeling Components



EMDR^2: Other Implementation Details

- Framework: PyTorch
  - Implemented using “*megatron-lm*” toolkit

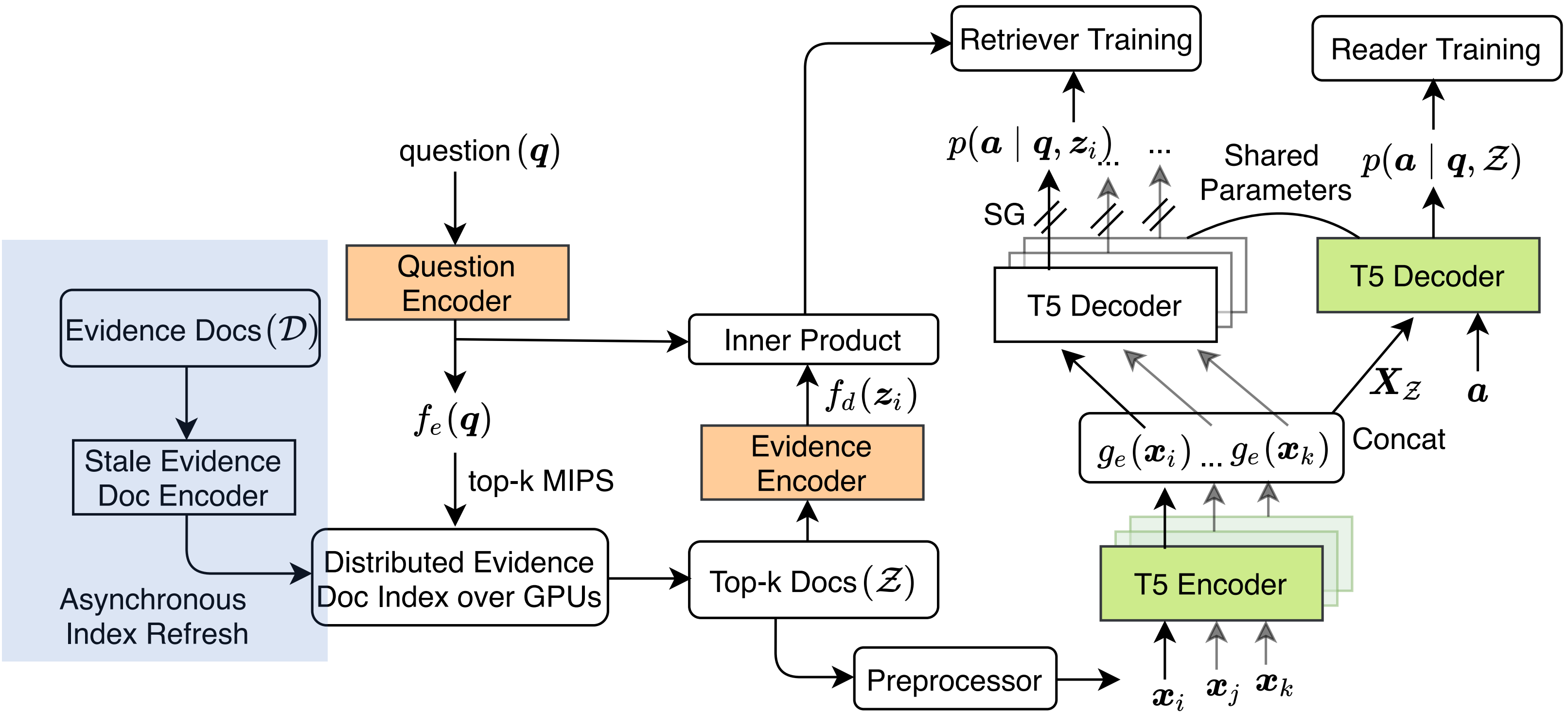
- **Compute**: 16 A100 GPUs, each with 40GB RAM

- 8 GPUs for model training (1st process group)

- 8 GPUs for asynchronous evidence indexing (2nd process group)
  - required because evidence embeddings get stale
  - performed every 500 training steps

- All 16 GPUs for top-K document retrieval (3rd process group).

EMDR²: Asynchronous Evidence Indexing



Comparison of Open-Domain QA Approaches

			Reader and Retriever Training			
Model	*Multi-Doc Reader*	*Retriever Adaptation*	*Disjoint*	*End-to-End*	*Multi-Step*	*Unsupervised Retriever*
REALM (Guu *et al.*, 2020)		✓		✓		✓
DPR (Karpukhin *et al.*, 2020)			✓			
RAG (Lewis *et al.*, 2020b)		✓		✓		
FiD (Izacard and Grave, 2021b)	✓		✓			
FiD-KD (Izacard and Grave, 2021a)	✓	✓			✓	
EMDR^2 (Our Approach)	✓	✓		✓		✓

Experimental Setting

- Base configuration of T5 and BERT (768 dim hidden size)
- **Total parameters**: 440M (T5-220M + Retriever-220M)
- **Batch Size**: 64
- **Top-K Documents**: 50
- **Evaluation**: Exact Match (EM)

EMDR^2: Unsupervised Pre-Training

- Helps to improve initial retrieval accuracy.
- **Corpus**: sentences containing named entities from Wikipedia.
- **Masked Salient Spans (MSS)**
  - Question: sentence with named entities masked
  - Answer: named entities
- Train for 82K steps with ICT initialized retriever.

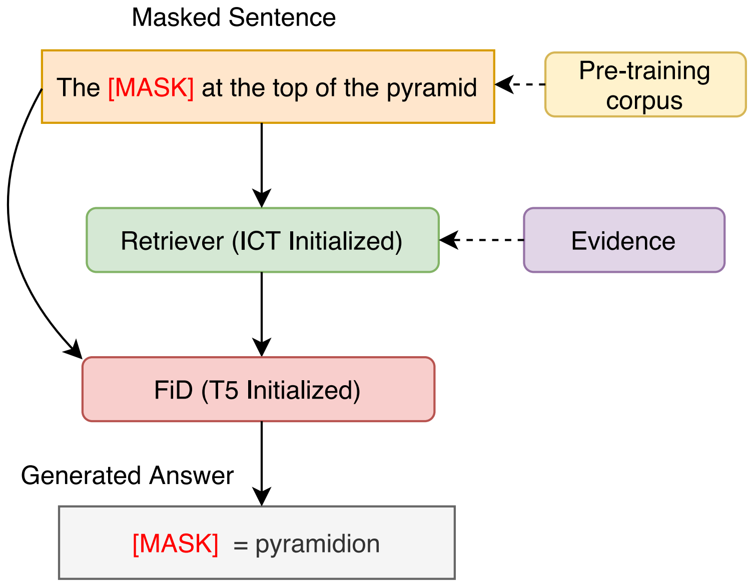


Figure derived from “Guu et al., 2020. REALM: Retrieval-Augmented Language Model Pre-Training”

Datasets

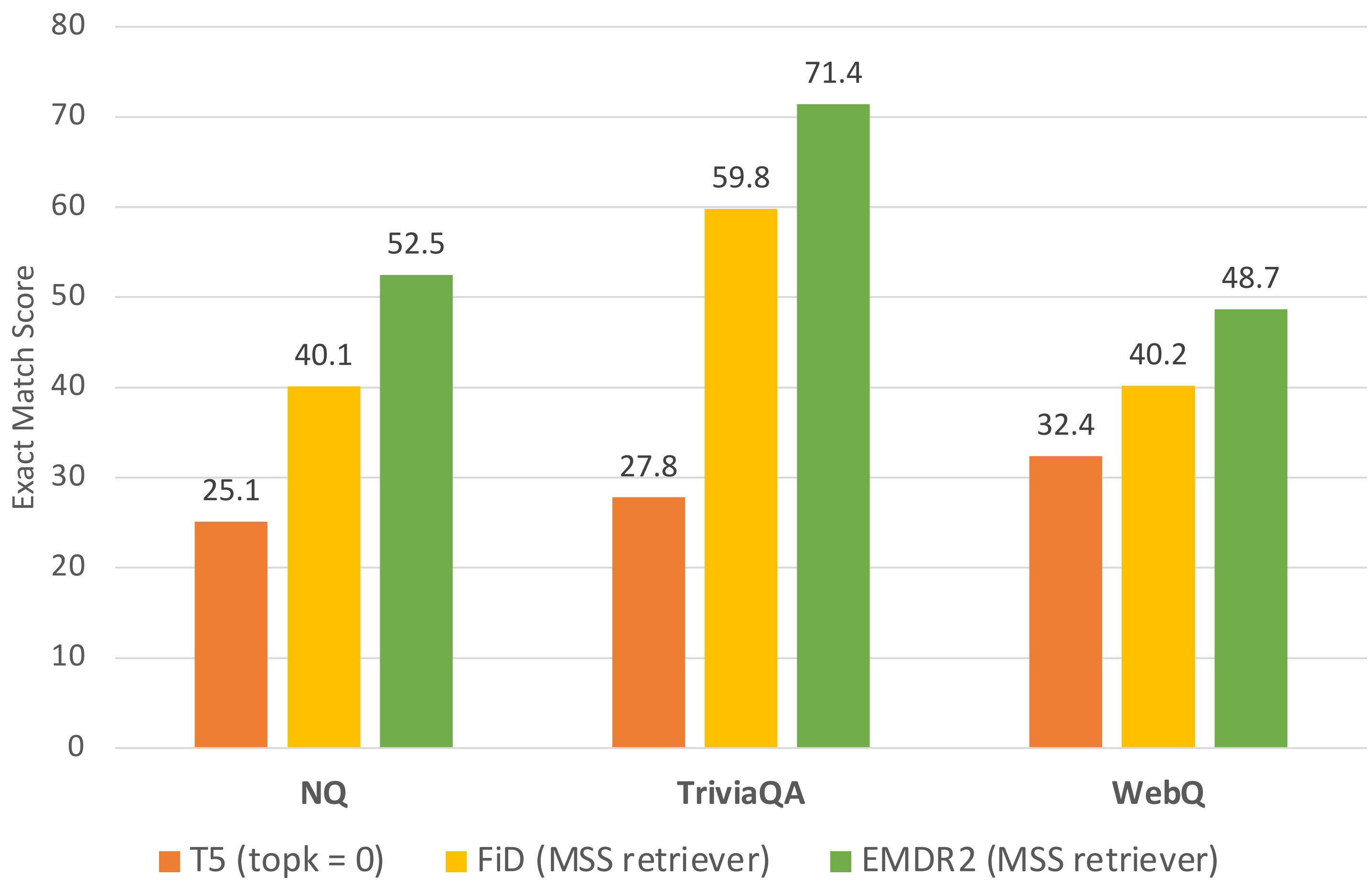
Dataset	Train	Dev	Test
WebQuestions (**WebQ**)	3,417	361	2,032
Natural Questions (**NQ**)	79,168	8,757	3,610
TriviaQA	78,785	8,837	11,313

WebQ: Questions were collected using Google Suggest API.
Freebase IDs in answers are replaced by entity names.

NQ: Real questions asked by users in Google. We use the subset of short answers.

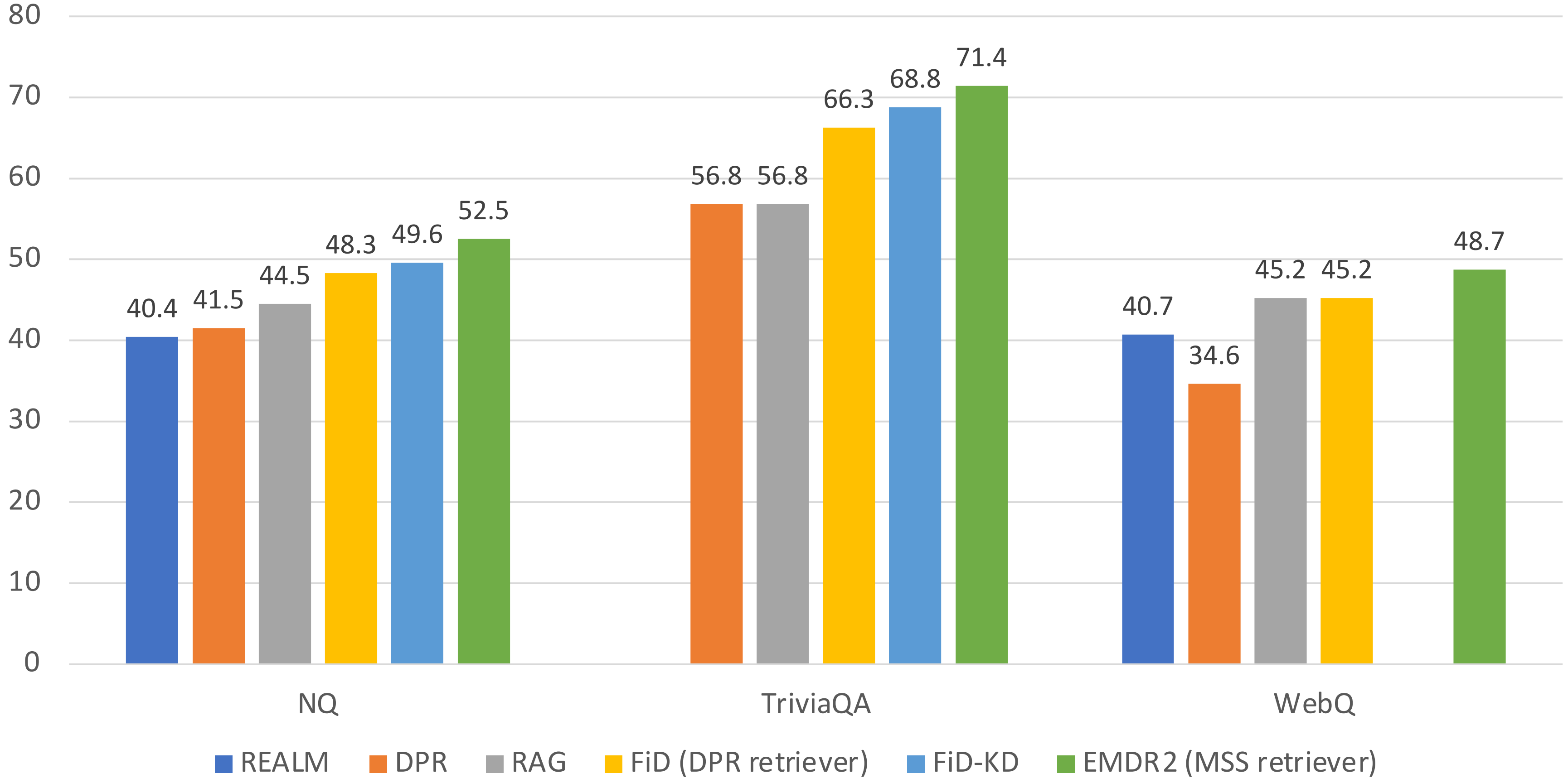
TriviaQA: Collection of trivia question-answer pairs collected from the web.

Results: EMDR^2 Training



End-to-end training provides good performance gains over FiD with MSS retriever

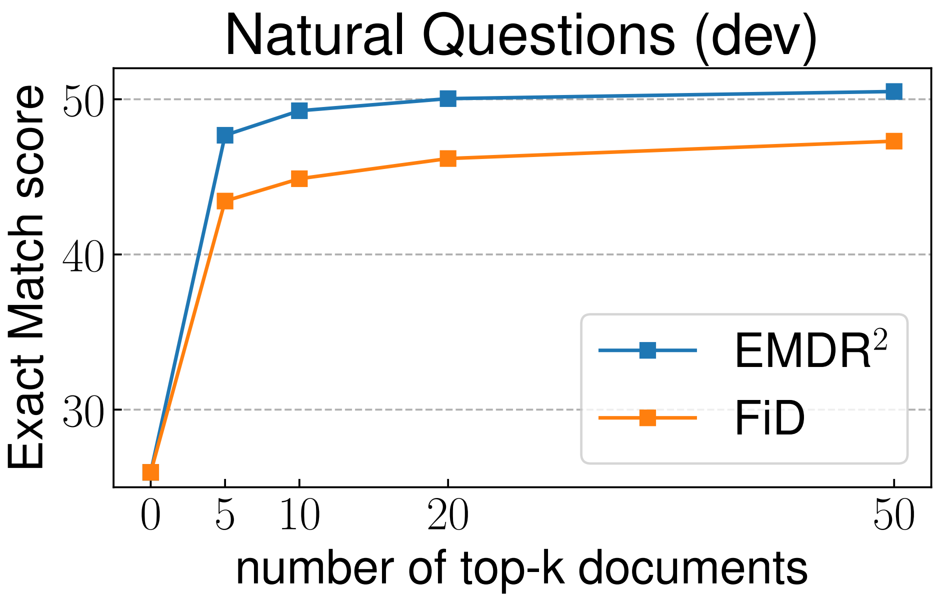
Comparison with Other Approaches

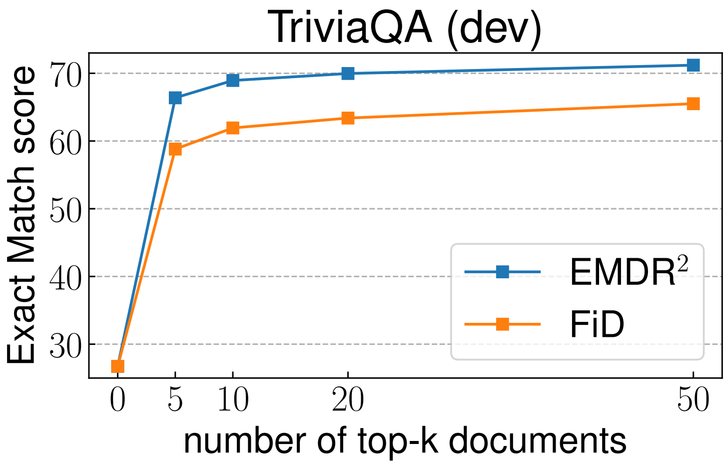


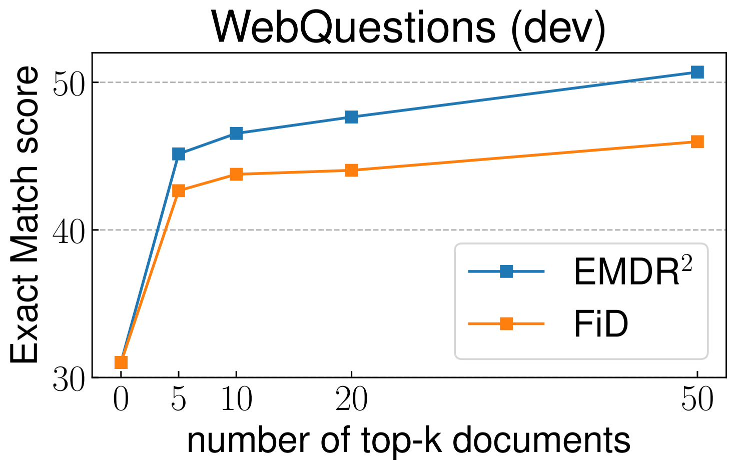
New SOTA Results of EMDR^2

2-3 EM points gain over FiD-KD

Analysis: Effect of the Number of Top-k Documents







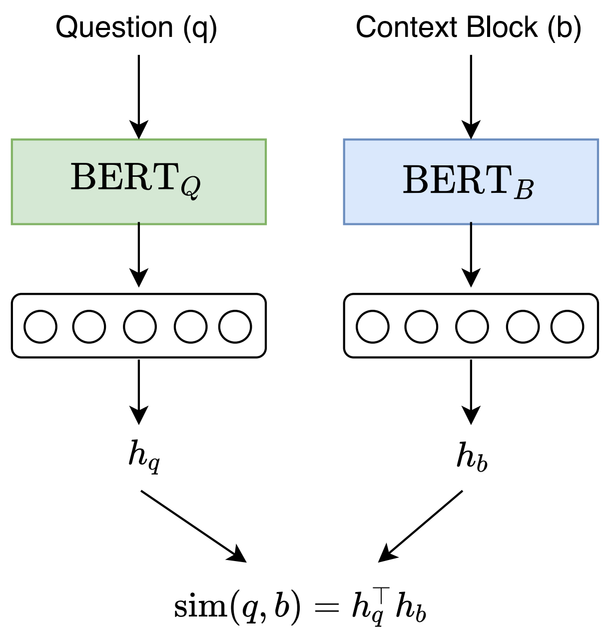
Effect of Retriever Initialization

Approaches compared:

1. Masked Salient Span (MSS) pre-training
2. Dense Passage Retrieval (DPR) training
3. MSS pre-training + DPR training

Review: Dense Passage Retrieval (DPR)

- Dual-encoder model
- Train from **supervised** question-context pairs



$D = q_i,$ Question

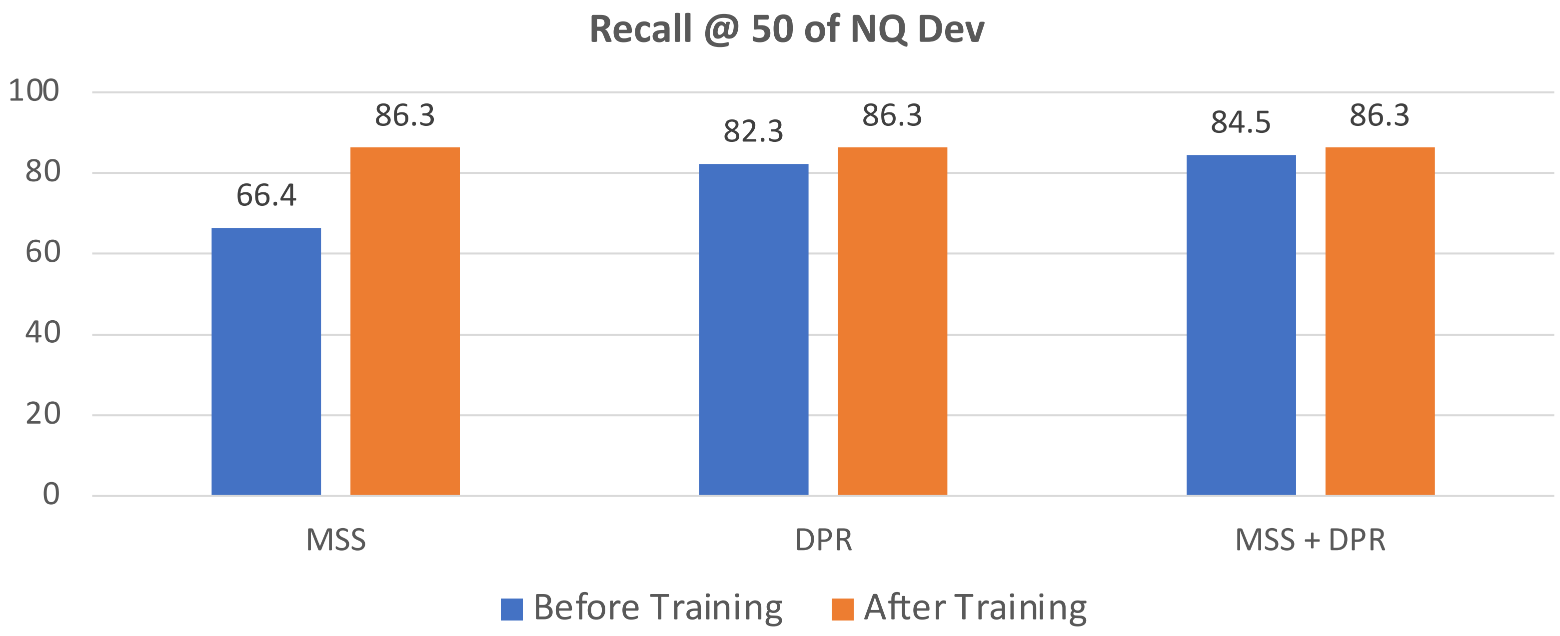
$p_i^+,$ One Positive Example

$p_{i,j}^-$ Set of Negative Examples

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

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

Effect of Retriever Initialization



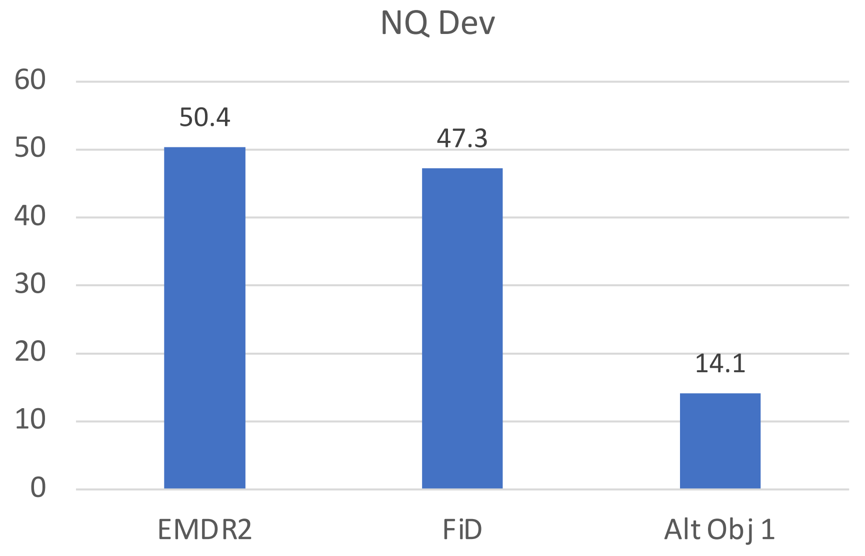
- MSS, DPR, and MSS + DPR retriever initialization results in the same final retrieval accuracy.
- Retriever training by DPR may not be essential for open-domain QA.

Alternative Training Objective 1

$$p(\mathcal{Z} \mid \boldsymbol{q}; \Phi) = \prod_{k=1}^{K} p(\boldsymbol{z}_k \mid \boldsymbol{q}; \Phi)$$

No feedback from FiD reader to retriever

$$\mathcal{L}_{\text{alt-1}} = \underbrace{\log p(\boldsymbol{a} \mid \boldsymbol{q}, \mathcal{Z}; \Theta)}_{\text{FiD reader}} + \underbrace{\sum_{k=1}^{K} \log p(\boldsymbol{z}_k \mid \boldsymbol{q}, \mathcal{Z}; \Phi)}_{\text{retriever}}$$

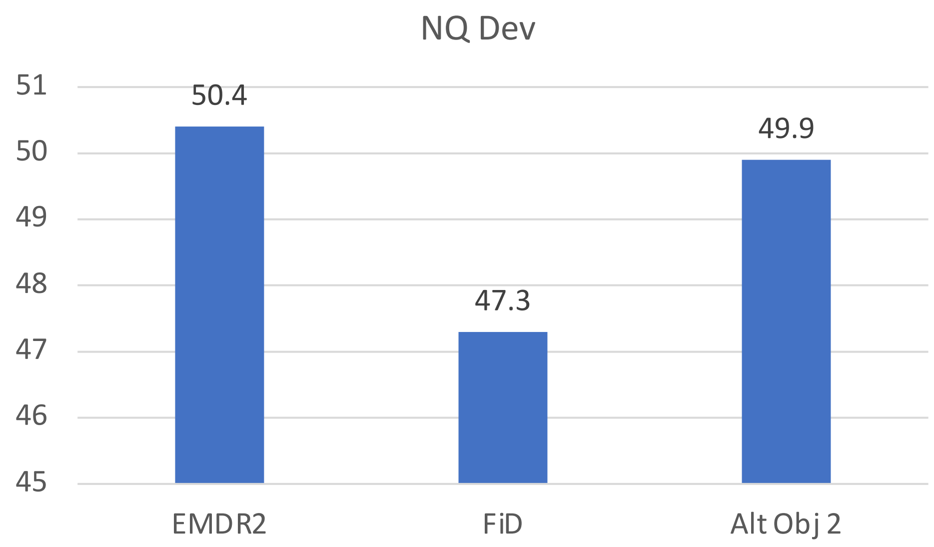


- Poor performance of Alt. Obj.
- Retriever gets stuck in a bad local optimum

Alternative Training Objective 2

$$\tilde{p}(\boldsymbol{a} \mid \boldsymbol{q}, \boldsymbol{z}_k; \Theta) = \frac{p(\boldsymbol{a} \mid \boldsymbol{q}, \boldsymbol{z}_k; \Theta)}{\sum_{j=1}^{K} p(\boldsymbol{a} \mid \boldsymbol{q}, \boldsymbol{z}_j; \Theta)}$$

$$\mathcal{L}_{\text{alt-2}} = \log p(\boldsymbol{a} \mid \boldsymbol{q}, \mathcal{Z}; \Theta) + \mathbb{KL}(\mathbb{SG}\,(\tilde{p}(\boldsymbol{a} \mid \boldsymbol{q}, \boldsymbol{z}_k; \Theta)) \mid\mid p(\boldsymbol{z}_k \mid \boldsymbol{q}, \mathcal{Z}; \Phi))$$



This objective offers improvement over the FiD model

Future Work

1. Application to knowledge-grounded dialogue generation.

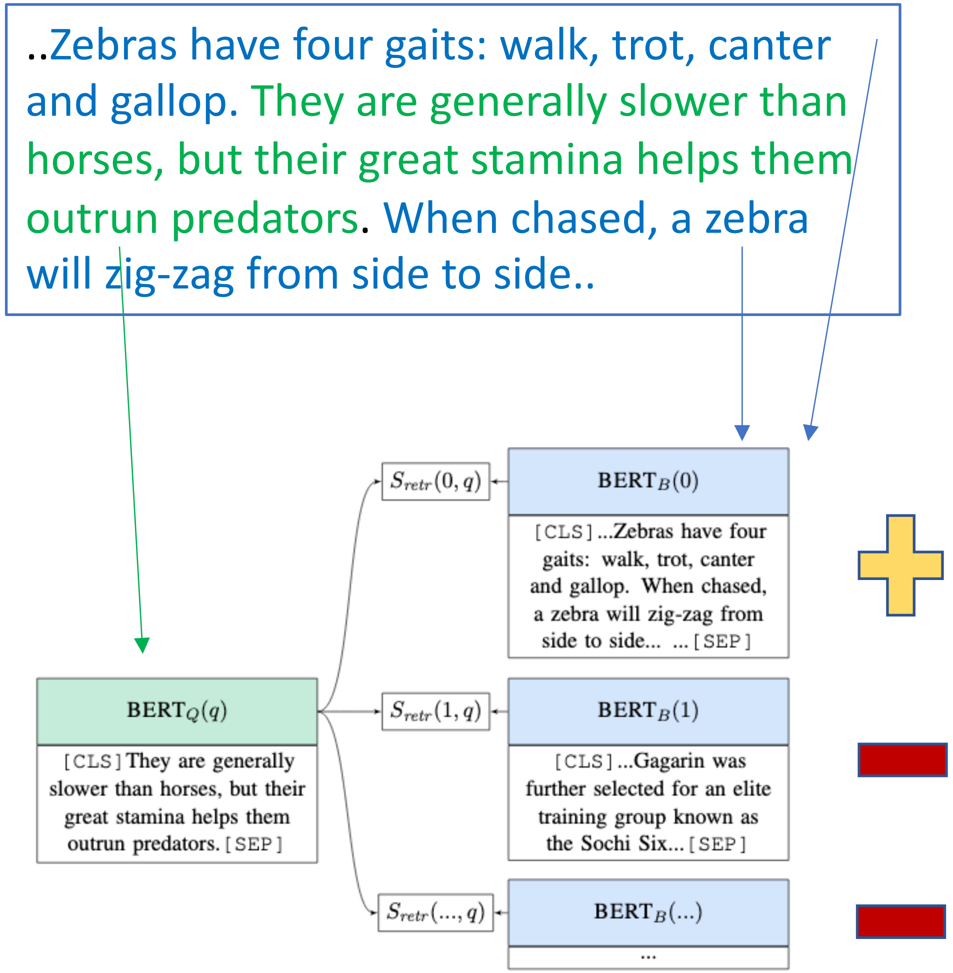
2. Multilingual open-domain question answering.

Thank you and Questions!

Code to be released at: https://github.com/DevSinghSachan/emdr2

Extra Slides: Dual Encoder Initialization by ICT

- Inverse Cloze Task (ICT)
- Sample a sentence from a paragraph.
- Sentence can be considered a *pseudo-query.*
- Remaining sentences can be considered as a *pseudo-context.*
- **Unsupervised** - can use all Wikipedia to train the model.



Things which didn't work

- FiD: concatenating the top-K documents together and increasing the position embeddings to 12000.
- FiD: concatenating the top-K documents together and introducing K segment embeddings.
- Asynchronous embedding updates with 250 steps not much improvements.