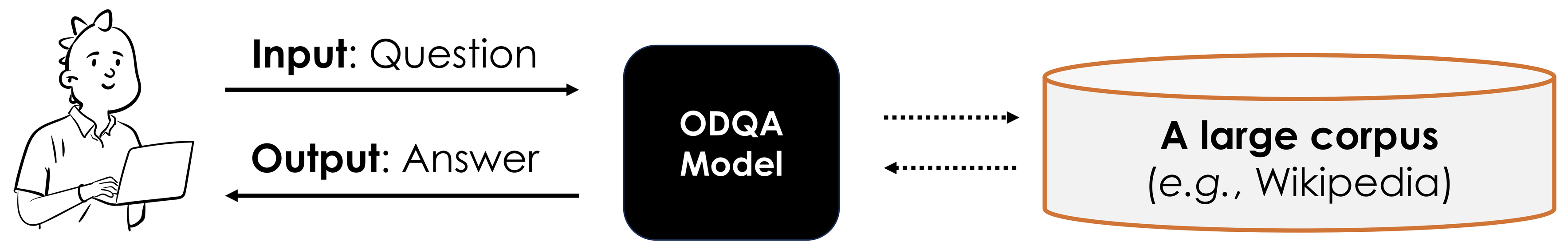
Multi-Granularity Guided Fusion-in-Decoder

Eunseong Choi, Hyeri Lee, Jongwuk Lee

Sungkyunkwan University (SKKU), Republic of Korea

NAACL 2024

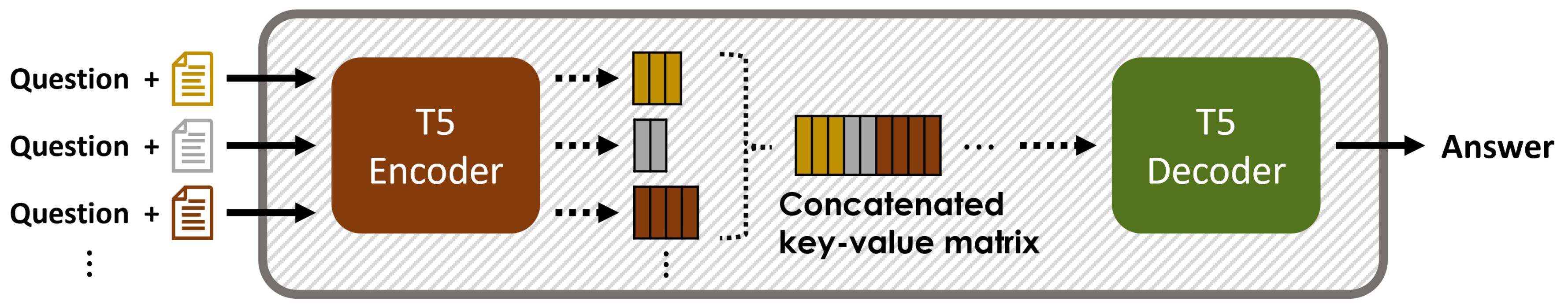
Task: Open-Domain Question Answering



Open-**D**omain **Q**uestion **A**nswering (**ODQA**) requires **answering factual questions** with reference to **a large corpus**.

- ✓ One of the **Knowledge-Intensive Language Tasks** (KILTs) that a human is unlikely to perform without access to an external knowledge source
- ✓ Impractical to examine every single document in the corpus
  → **Retrieve-then-Read pipeline**

Fusion-in-Decoder (FiD)



Fusion-in-Decoder (FiD) is a notable method which concatenates multiple encoder outputs and feeds them as a key-value matrix to the decoder, which we define as a **Multi-document Reader**.

- ✓ Aggregating evidence across multiple passages

Takeaways

- ✓ **Multi-granularity-based evidentiality** for the FiD architecture
- ✓ LLM generated **pseudo-labels for supportive passages**
  - filtering out spurious ones that contain the answer span
- ✓ **Reusing multi-granularity contexts for accuracy and efficiency**
  - an anchor vector for the decoder
  - threshold-based passage pruning

Challenges in Multi-document Readers

Question: who played in the most world series games
Answer: the New York Yankees
Retrieved passage: **World Series** television ratings The highest average rating for an entire **World Series** is tied between the 1978 Series featuring **the New York Yankees** and **Los Angeles Dodgers** and the 1980 Series featuring the **Philadelphia** ...
Answer span: True / **Supportive**: False

Question: when did the first manned space craft land on the moon
Answer: 20 July 1969
Evidence passage: ... includes both **manned** and unmanned (robotic) missions. The first human-made object to reach the surface of the Moon was the Soviet Union's Luna 2 mission, on **13 September 1959**. **The United States' Apollo 11 was the first manned mission to land on the Moon, on 20 July 1969**. There have been six **manned** U.S. landings (between **1969** and **1972**) ... unmanned landings, from **22 August 1976** until **14 December 2013**.
Prediction w/o *multi-granularity learning*: **13 September 1959**

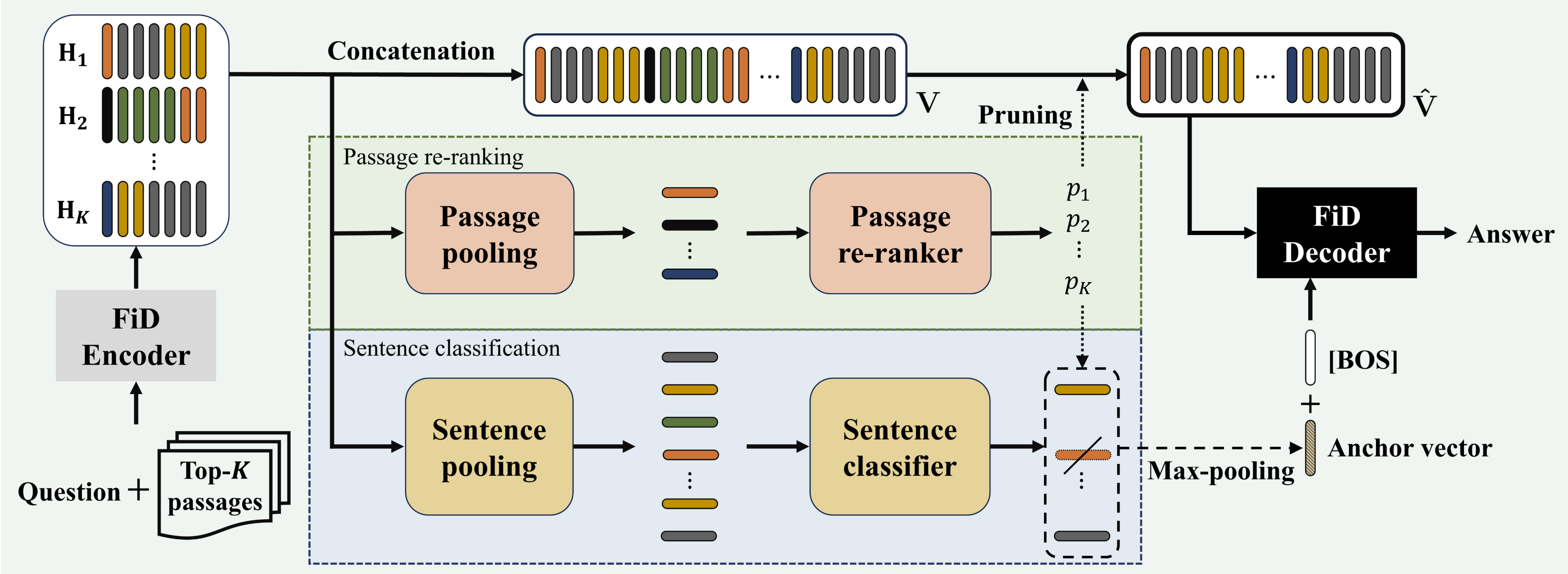
Challenge ①
Discriminate supportive passages among the retrieved result
- ✓ Even the passage containing an answer span can be spurious.

Potentially misleading | Supportive sentence | Answer span

Challenge ②
There are distractions that potentially mislead the reader.

→ **We propose to discern evidence across multiple levels of granularity, *i.e.*, passages and sentences.**

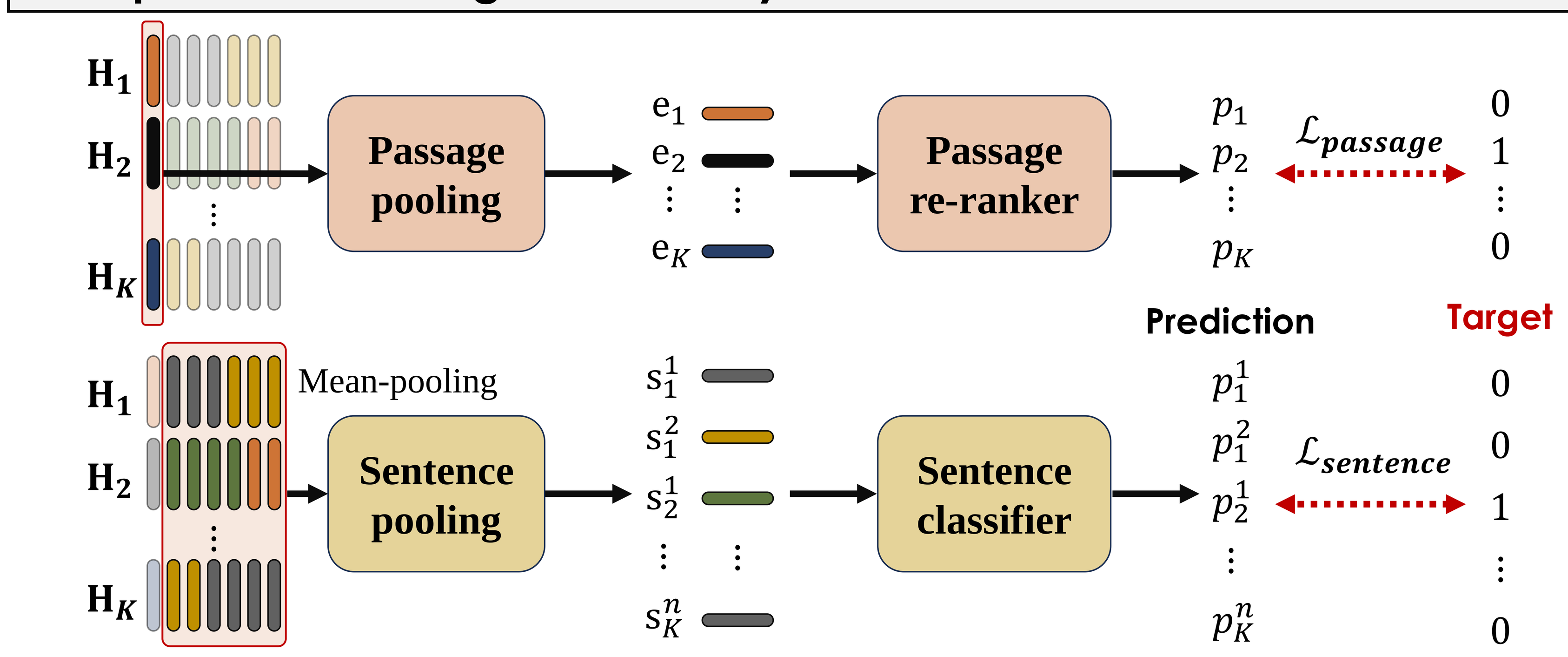
Multi-**G**ranularity Guided **F**usion-**i**n-**D**ecoder (**MGFiD**)



MGFiD framework incorporates 1) **Answer Generation** and 2) **Learning & Reusing Multi-Granularity Contexts.**

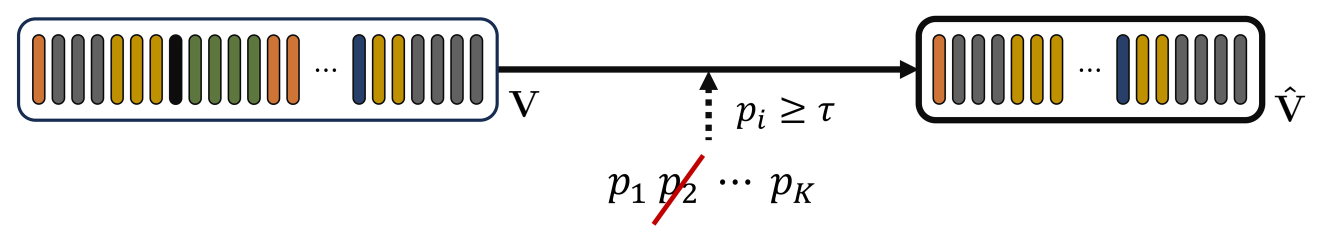
1. **Answer Generation**
   - ✓ Adopting standard FiD architecture
     - FiD-decoder takes the concatenated matrix $\mathbf{V} \in \mathbb{R}^{(K \times L) \times d}$ as the key-value matrix and generates an answer.
     - K: number of candidate passages / L: maximum sequence length / d: hidden dimension
2. **Learning & Reusing Multi-Granularity Evidence**
   - ✓ **Passage re-ranking to identify coarse-grained evidence**
   - ✓ **Sentence classification for fine-grained evidence**

Leveraging the ranking capabilities of Large Language Models, MGFiD uses pseudo-labels generated by LLMs.



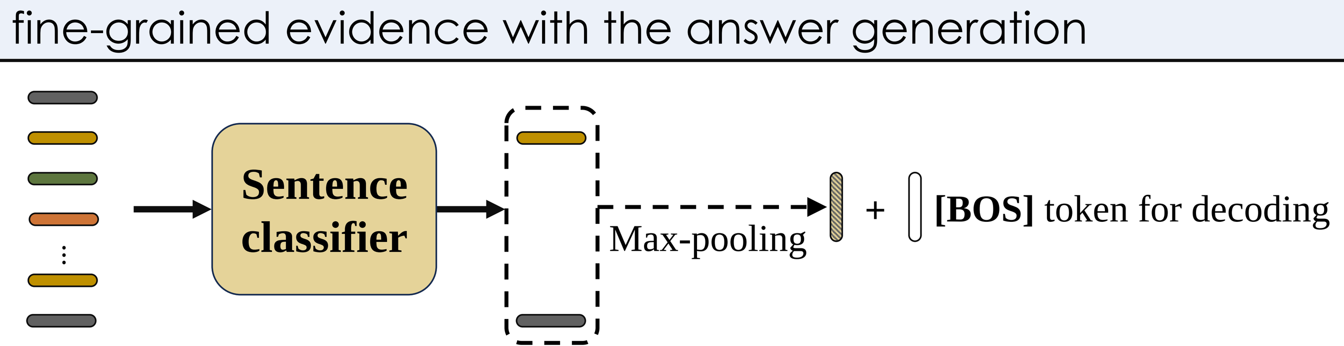
Threshold-based passage pruning

- ✓ Mitigating inefficiency from the huge key-value matrix



Anchor vector

- ✓ Aligning fine-grained evidence with the answer generation



Experimental Results

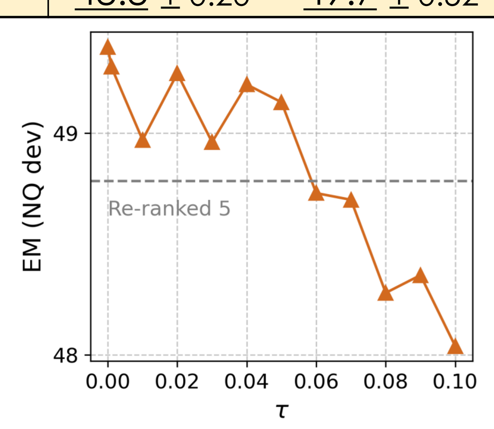
Effectiveness on Question Answering datasets

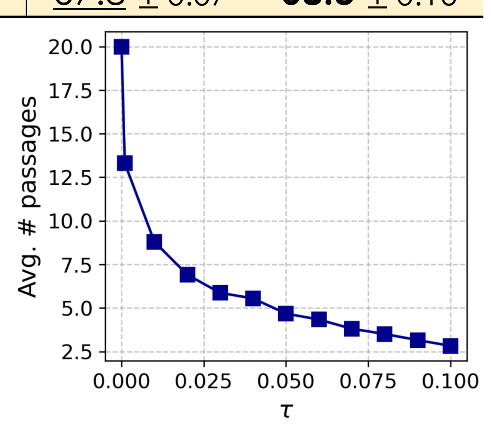
- ✓ MGFiD significantly improves the effectiveness of both datasets using the same retriever and the same number of passages.
- ✓ Pruned MGFiD outperforms baselines using **24%~39% passages in the decoder.**

Model	Multi-task learning	Ret.	Avg. # psgs in Decoder	NQ (EM)		TQA (EM)	
				Dev	Test	Dev	Test
FiD-KD	-	FiD-KD	20	47.8 ±0.16	48.4 ±0.31	67.4 ±0.12	67.6 ±0.25
FiD-KD	-		100	49.1	50.1	-	-
EvidentialityQA	O		20	48.0 ±0.20	49.0 ±0.39	n/a	n/a
RFiD	O		20	48.6 ±0.29	49.4 ±0.53	67.8 ±0.12	68.1 ±0.20
RFiD	O		100	49.2	50.4	-	-
MGFiD	O	FiD-KD	20	**49.0** ±0.21	**50.1** ±0.33	**68.0** ±0.09	**68.3** ±0.23
Pruned MGFiD	O		4.8 / 7.7	48.8 ±0.20	49.7 ±0.52	67.8 ±0.07	**68.3** ±0.16

Effectiveness varying the pruning threshold τ

- ✓ Increasing τ to 0.05 results in a small performance drop even if the number of passages drops drastically below 5.





QA performance with different passage labels

Passage label	NQ dev (EM)	# positives	TQA dev (EM)	# positives
-	47.8 ±0.16	-	67.4 ±0.12	-
Answer span	48.5 ±0.21	4.5	67.7 ±0.16	8.9
ChatGPT	**48.9** ±0.13	4.0	67.7 ±0.20	8.3
MythoMax	48.8 ±0.23	2.8	**67.8** ±0.18	6.5

Ablation studies

- ✓ **Listwise loss function for passage re-ranking achieves higher accuracy** than pointwise loss function on both datasets.
- ✓ **With the anchor vector, the effectiveness improved by 0.5%p** on the NQ dataset.

$\mathcal{L}_{passage}$	$\mathcal{L}_{sentence}$	$\mathbf{e}_{anchor}$	τ	NQ dev	TQA dev
listwise	✓	✓	0.05	49.1	67.7
listwise	✓	✓	*top-5*	48.8	-
listwise	✓	✓	X	**49.4**	**67.9**
listwise	✓	X	X	48.9	**67.9**
X	✓	X	X	48.1	**67.9**
listwise	X	X	X	48.8	67.8
pointwise	X	X	X	48.3	67.6
X	X	X	X	47.8	67.5

Check out our paper, details on MGFiD, and more experiments!

contact info.
eunseong@skku.edu

Paper | Code