NAACL 2024

Multi-Granularity Guided Fusion-in-Decoder

Eunseong Choi, Hyeri Lee, Jongwuk Lee

Sungkyunkwan University (SKKU), Republic of Korea

Introduction

Task: Open-Domain Question Answering

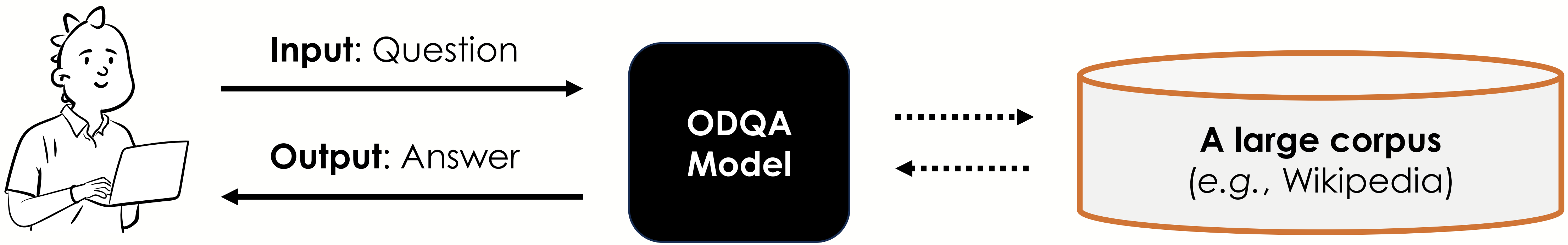
Retrieval-Augmented Generation (RAG)

Challenges in Multi-document Readers

Key Contribution

Task: Open-Domain Question Answering

➢ **Open-Domain Question Answering (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**

Retrieval-Augmented Generation (RAG)

- **RAG, employing the "Retrieve-then-Read" pipeline, uses retrieved results as input for the generative model.**
  - To deliver high performance,
    - ✓ How to effectively retrieve evident documents?
    - ✓ **How to effectively read retrieved results?**

Retrieval-Augmented Generation (RAG)

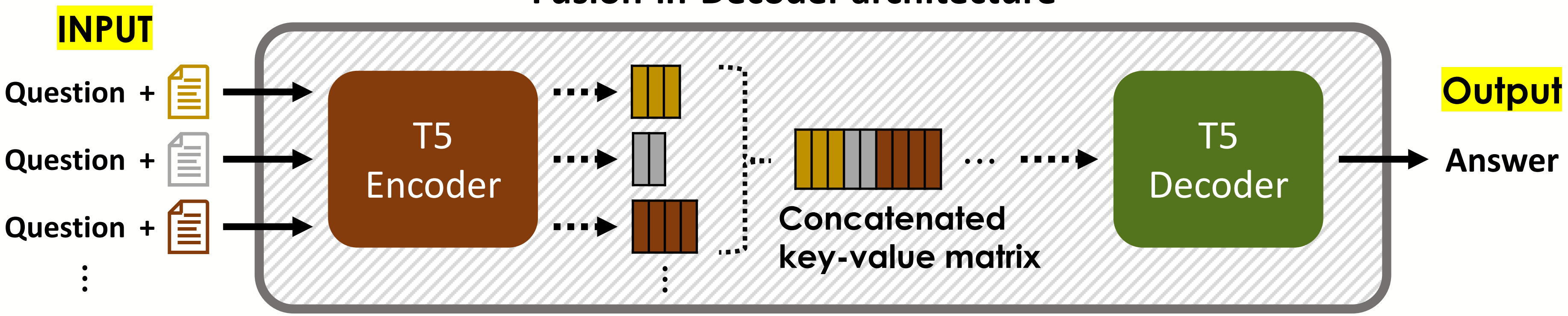
- **RAG, employing the "Retrieve-then-Read" pipeline, uses retrieved results as input for the generative model.**
  - To deliver high performance,
    - ✓ How to effectively retrieve evident documents?
    - ✓ **How to effectively read retrieved results?**

Retrieval-Augmented Generation (RAG)

- **RAG, employing the "Retrieve-then-Read" pipeline, uses retrieved results as input for the generative model.**
  - To deliver high performance,
    - ✓ How to effectively retrieve evident documents?
    - ✓ **How to effectively read retrieved results?**

- **Fusion-in-Decoder (FiD) is a notable method which concatenates multiple encoder outputs and feeds it as a key-value matrix in the decoder.**

Fusion-in-Decoder architecture

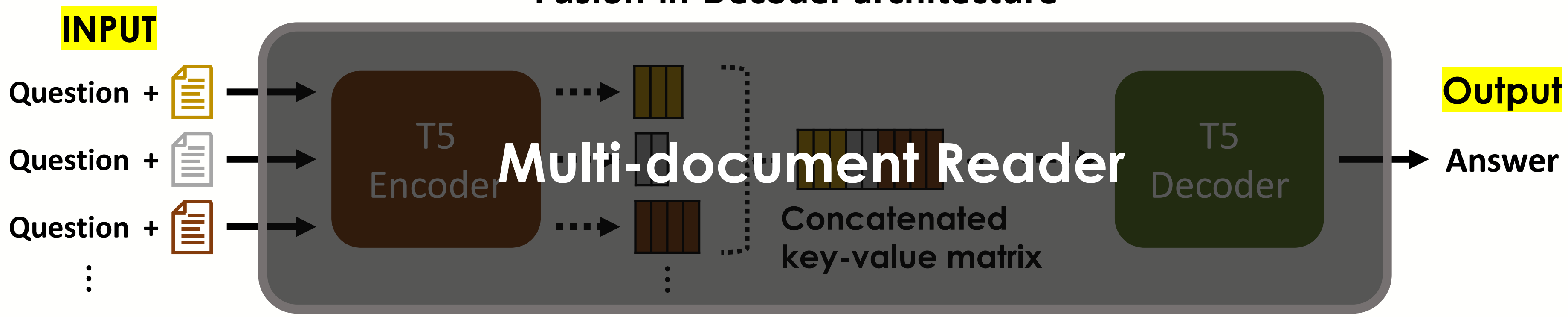


Retrieval-Augmented Generation (RAG)

- **RAG, employing the "Retrieve-then-Read" pipeline, uses retrieved results as input for the generative model.**
  - To deliver high performance,
    - ✓ How to effectively retrieve evident documents?
    - ✓ **How to effectively read retrieved results?**

- **Fusion-in-Decoder (FiD) is a notable method which concatenates multiple encoder outputs and feeds it as a key-value matrix in the decoder.**

Fusion-in-Decoder architecture



Challenges in Multi-document Readers ①

- **It is required to discriminate supportive passages among the retrieved result.**
  - Existing works adopt multi-task learning to enhance the ability to discriminate.
    - e.g., $\mathcal{L} = \mathcal{L}_{QA} + \mathcal{L}_{classifiaction}$
  - As Open QA settings do not include annotations for gold passages, they **usually rely on a silver standard**, where the **passages containing the answer span are used for** $\mathcal{L}_{classification}$.

- **However, some passages are not supportive although contain the answer span.**

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**

Potentially misleading
Answer span

Challenges in Multi-document Readers ②

➢ **Although it successfully identifies an evident passage, there still exist distractors that may confuse the reader.**

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**

Supportive
Potentially misleading
Answer span

Challenges in Multi-document Readers ②

- **Although it successfully identifies an evident passage, there still exist distractors that may confuse the reader.**

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

Supportive
Potentially misleading
Answer span

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

Key Contribution

1. We introduce evidentiality to the FiD architecture using multi-granularity contexts.

2. We utilize LLMs to generate pseudo-labels for supportive passages, thereby filtering out spurious ones that contain the answer span.

3. We further improve accuracy and efficiency by reusing multi-granularity contexts, incorporating an anchor vector in the decoder, and employing threshold-based passage pruning.

Proposed Method

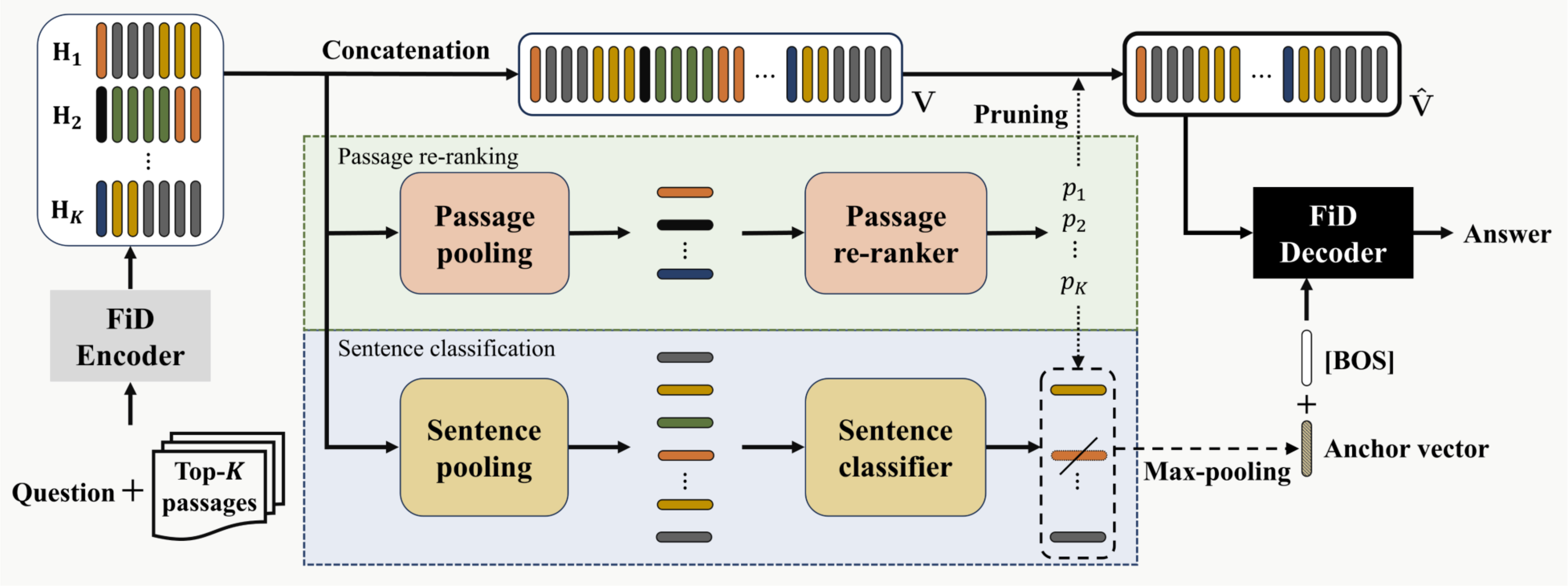
Overall Architecture

Answer Generation

Learning & Reusing Multi-Granularity Contexts

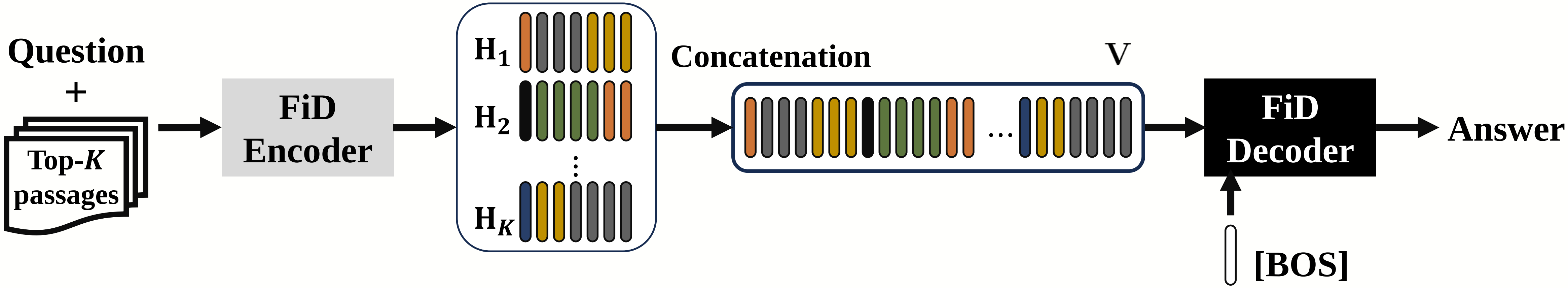
Overall Architecture

- **M**ulti-**G**ranularity Guided **F**usion-**i**n-**D**ecoder (**MGFiD**) incorporates multi-task learning for answer generation.
  1) **Passage re-ranking** to identify coarse-grained evidence
  2) **Sentence classification** for fine-grained evidence



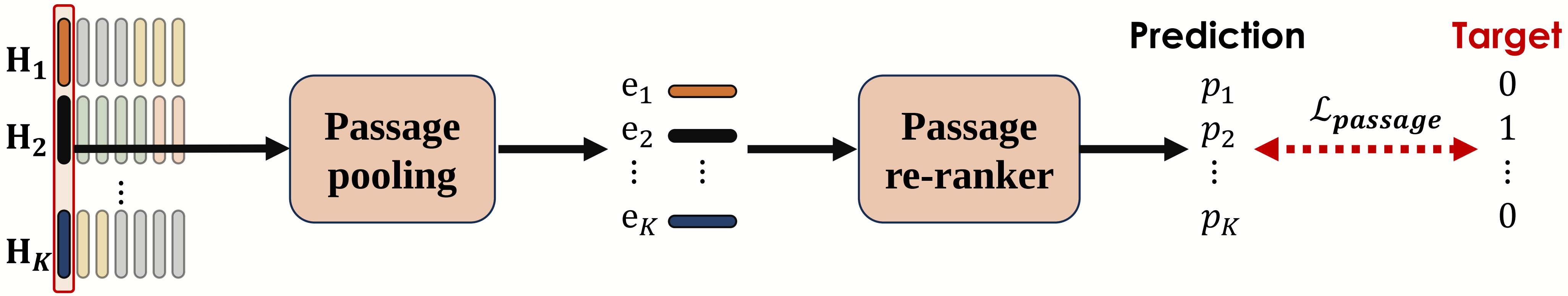
Answer Generation

➢ **MGFiD adopts the standard FiD architecture to generate the answer.**
- **FiD-encoder** outputs K token embeddings $\mathbf{H}_i \in \mathbb{R}^{L \times d}$.
  - L: maximum sequence length, d: hidden dimension
- **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 auto-regressively.

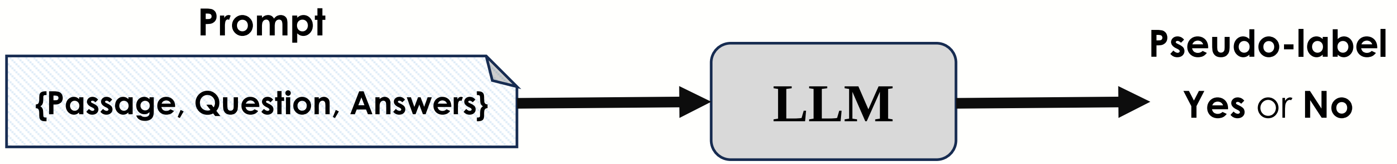


Learning Multi-Granularity Contexts ①

1. **Identifying coarse-grained evidence through *passage re-ranking***
   - **Passage pooling** takes the first token embedding from the question-passage pair $\mathbf{H}_i$ and obtains an evident embedding e_i through a projection layer.
   - **Passage re-ranker** produces the probability that each passage is supportive.

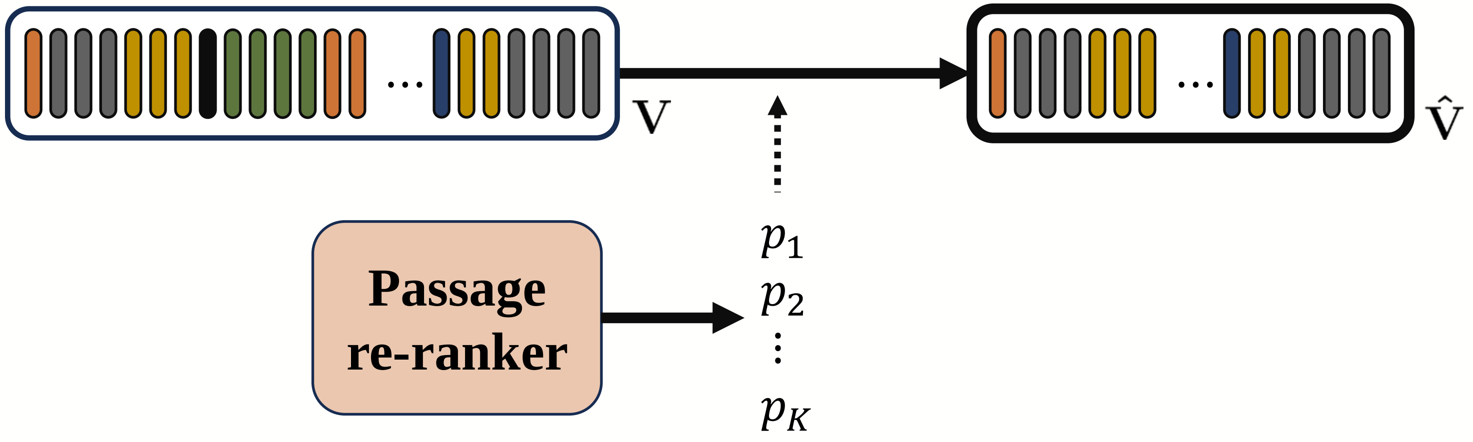


   - Leveraging the ranking capabilities of Large Language Models (LLMs), we use **pseudo-labels generated by LLMs according to the evidentiality of passages.**



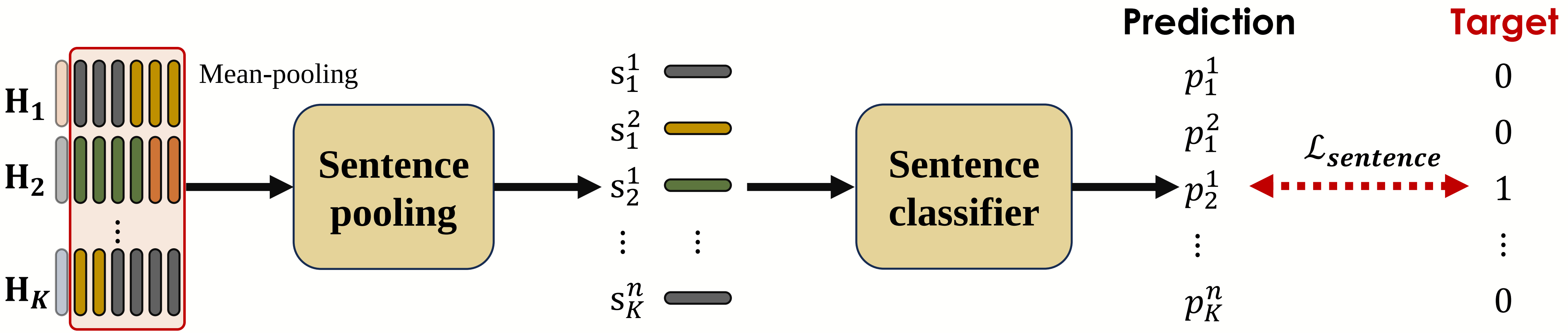
Reusing Multi-Granularity Contexts ①

- **Passage pruning** discards passages below a threshold τ to enhance efficiency.
  - Inefficiency from the huge key-value matrix is a major drawback of FiD architecture.
  - By reusing the probabilities, MGFiD can prune the spurious passages based on a threshold.



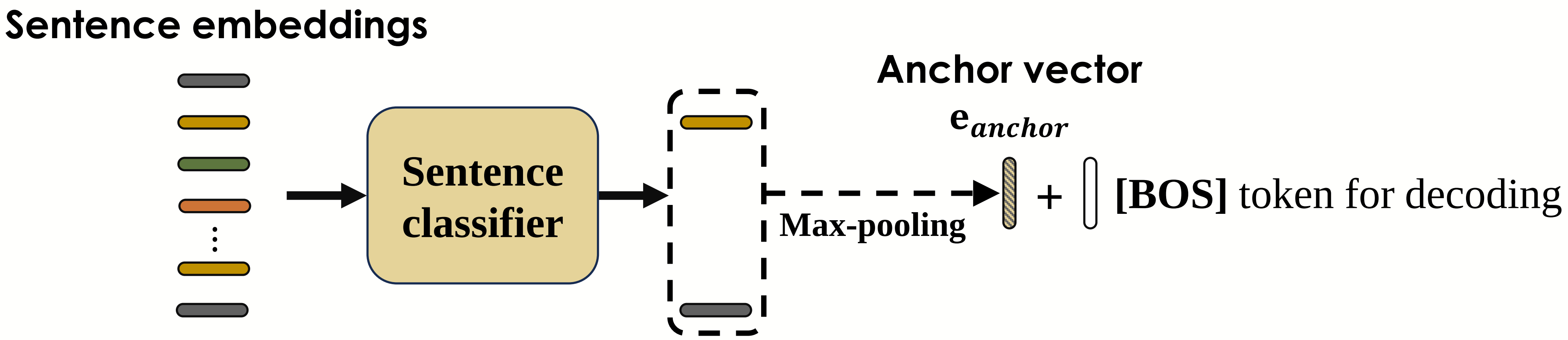
Learning Multi-Granularity Contexts ②

2. **Identifying fine-grained evidence through *sentence classification***
   - **Sentence pooling** obtains sentence embeddings through **mean-pooling of token embeddings**.
   - As the passages have been assessed by the LLM, sentences containing the answer span are used as the target.



Reusing Multi-Granularity Contexts ②

- **Anchor vector** aligns the multi-task of identifying evidence with the answer generation.
  - Although multi-task learning highlights evidential contexts, **how the decoder leverages this highlighted information remains unexplored.**



Experiments

Experimental Setup

Baselines

Experimental Results

Experimental Setup

- **We conduct experiments on Natural Questions (NQ) and TriviaQA (TQA).**

Dataset	Train	Dev	Test	R@20	Avg. # positives/query
Natural Questions (NQ)	79,168	8,757	3,610	0.87	4.5
Trivia QA (TQA)	78,785	8,837	11,313	0.86	8.9

  - **Recall@20 (R@20)** and **Avg. # positives/query** evaluate the passages containing answers among the retrieval results on train dataset of each.

- **Retriever**
  - Dense Passage Retriever (DPR)
    - A retriever that adopts bi-encoder architecture to evaluate the relevance between the query and the passage embeddings
  - **DPR-FiD-KD**
    - Dense Passage Retriever model trained through knowledge distillation from the FiD-KD

Baselines

➢ **We compare MGFiD with several FiD-based models with/without multi-task learning.**

1) **Models without multi-task learning**
   ① FiD
   ② FiD-KD
      - A difference between FiD and FiD-KD is the retriever used for the candidate passages.

2) **Models with multi-task learning**
   ① EvidentialityQA
   ② RFiD

- Except for the FiD, we compare all models under the same retrieval settings, where the retriever is FiD-KD and the number of passages (K) is set to 20.

Effectiveness on Question Answering datasets

➢ **MGFiD significantly improves the effectiveness of both datasets using the same retriever and the same number of passages.**

- We report the average results over five seeds and their standard deviations.
- Pruned MGFiD outperforms baselines using only 24%~39% passages in the decoder.

Model	Multi-task learning	Retriever	Avg. # psgs in Decoder	NQ (EM)		TQA (EM)	
				Dev	Test	Dev	Test
FiD	-	DPR	20	45.3 ± 0.31	46.3 ± 0.10	61.5 ± 0.12	62.1 ± 0.34
FiD-KD	-	DPR-FiD-KD	20	47.8 ± 0.16	48.4 ± 0.31	67.4 ± 0.12	67.6 ± 0.25
EvidentialityQA	O	DPR-FiD-KD	20	48.0 ± 0.20	49.0 ± 0.39	n/a	n/a
RFiD	O	DPR-FiD-KD	20	48.6 ± 0.29	49.4 ± 0.53	<u>67.8</u> ± 0.12	<u>68.1</u> ± 0.20
MGFiD	O	DPR-FiD-KD	20	**49.0** ± 0.21	**50.1** ± 0.33	**68.0** ± 0.09	**68.3** ± 0.23
Pruned MGFiD (τ=0.05)	O	DPR-FiD-KD	4.8 / 7.7	<u>48.8</u> ± 0.20	<u>49.7</u> ± 0.52	<u>67.8</u> ± 0.07	**68.3** ± 0.16

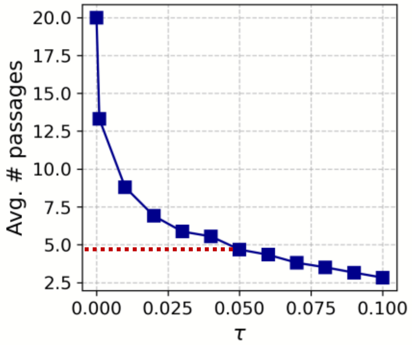
Effectiveness on Question Answering datasets

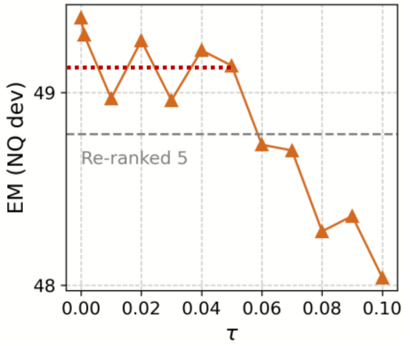
- **MGFiD significantly improves the effectiveness of both datasets using the same retriever and the same number of passages.**
  - Moreover, it achieves performance comparable to the baselines using 100 passages.

Model	Multi-task learning	Retriever	Avg. # psgs in Decoder	NQ (EM)		TQA (EM)	
				Dev	Test	Dev	Test
FiD	-	DPR	20	45.3 $\pm$0.31	46.3 $\pm$0.10	61.5 $\pm$0.12	62.1 $\pm$0.34
FiD-KD	-	DPR-FiD-KD	20	47.8 $\pm$0.16	48.4 $\pm$0.31	67.4 $\pm$0.12	67.6 $\pm$0.25
FiD-KD	-		100	49.1	50.1	-	-
EvidentialityQA	O		20	48.0 $\pm$0.20	49.0 $\pm$0.39	n/a	n/a
RFiD	O		20	48.6 $\pm$0.29	49.4 $\pm$0.53	<u>67.8</u> $\pm$0.12	<u>68.1</u> $\pm$0.20
RFiD	O		100	49.2	50.4	-	-
MGFiD	O	DPR-FiD-KD	20	**49.0** $\pm$0.21	**50.1** $\pm$0.33	**68.0** $\pm$0.09	**68.3** $\pm$0.23
Pruned MGFiD (τ=0.05)	O		4.8 / 7.7	<u>48.8</u> $\pm$0.20	<u>49.7</u> $\pm$0.52	<u>67.8</u> $\pm$0.07	**68.3** $\pm$0.16

In-depth Analysis

- **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.





- **Experiments with different passage labels**

Passage label	NQ dev (EM)	# positives	TQA dev (EM)	# positives
-	47.8 $\pm$ 0.16	-	67.4 $\pm$ 0.12	-
Answer span	48.5 $\pm$ 0.21	4.5	67.7 $\pm$ 0.16	8.9
ChatGPT	**48.9** $\pm$ 0.13	4.0	67.7 $\pm$ 0.20	8.3
MythoMax-13B	48.8 $\pm$ 0.23	2.8	**67.8** $\pm$ 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 (EM)	TQA dev (EM)
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

Conclusion

Conclusion

- We propose the **M**ulti-**G**ranularity Guided **F**usion-**i**n-**D**ecoder (**MGFiD**), a novel reader designed to manage evidence across multiple granularities.
  - MGFiD synergizes **coarse-level passage re-ranking** with **fine-level sentence classification**.

- MGFiD utilizes multi-granularity evidence by:
  - **pruning passages to enhance decoding efficiency.**
  - **constructing anchor vector to guide decoder toward significant evidence.**

- We demonstrate that **MGFiD improves upon the original Fusion-in-Decoder (FiD) by more than 3.5% and 1.0%** on Natural Questions and Trivia QA, respectively.

Thank you

Email: eunseong@skku.edu

Code: https://github.com/eunseongc/MGFiD