From Reading to Compressing: Exploring the Multi-document Reader for Prompt Compression

Eunseong Choi, Sunkyung Lee, Minjin Choi, June Park, Jongwuk Lee

Sungkyunkwan University (SKKU), Republic of Korea

Paper Code

EMNLP 2024

Takeaways

✓ Prompt compression connected with the Fusion-in-Decoder to capture the global context

✓ Aligning the question-answering task with prompt compression

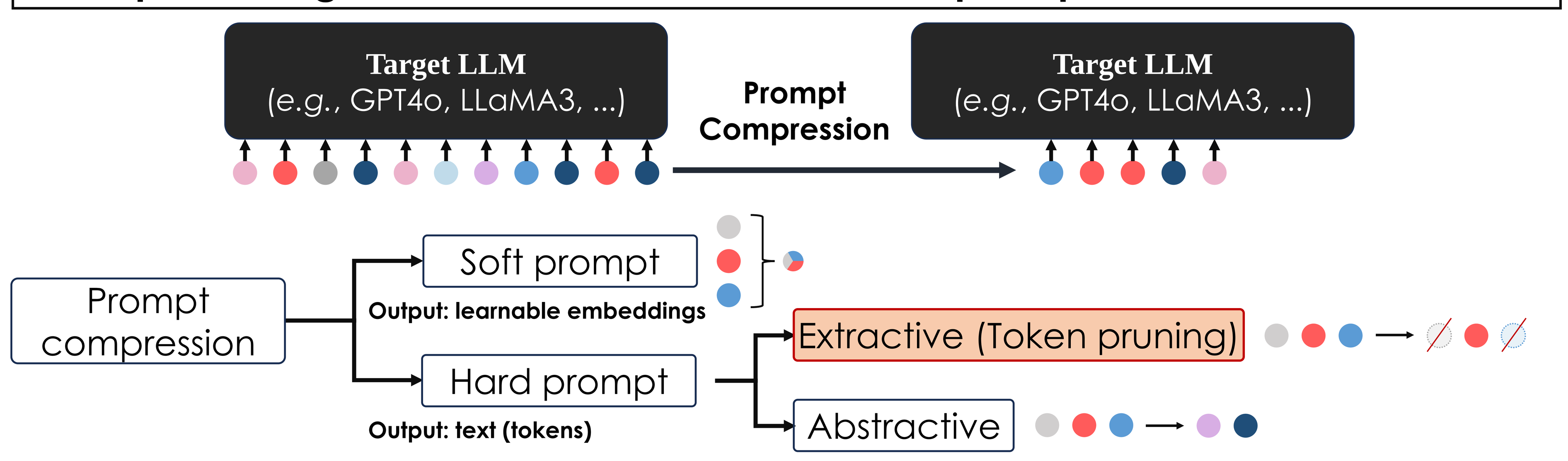
✓ The cross-attention scores trained to answer the question provide the effective approximation that the target LLM needs to focus on.

Token pruning for prompt compression

- (**vs. Soft prompt**) easily adaptable to any black-box LLMs: no need to train an LLM
- (**vs. Abstractive compression**) more efficient than abstractive compression, where auto-regressive generation is required

Prompt Compression

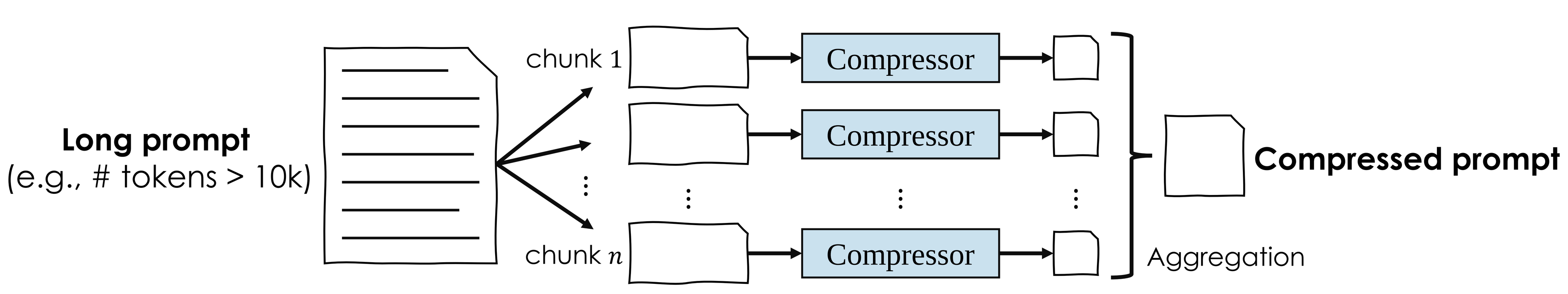
Prompt compression aims to reduce the computational overhead in LLMs, while preserving the essential information in the prompt.



Challenges in Token Pruning

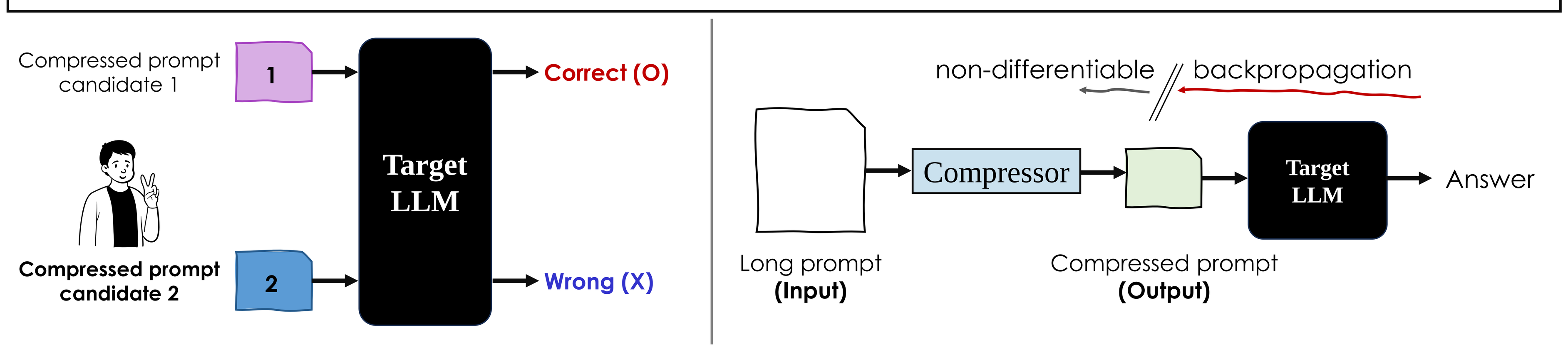
Challenge ① How to extract essential information based on the global context

- Existing work processes chunk by chunk, neglecting the global context.

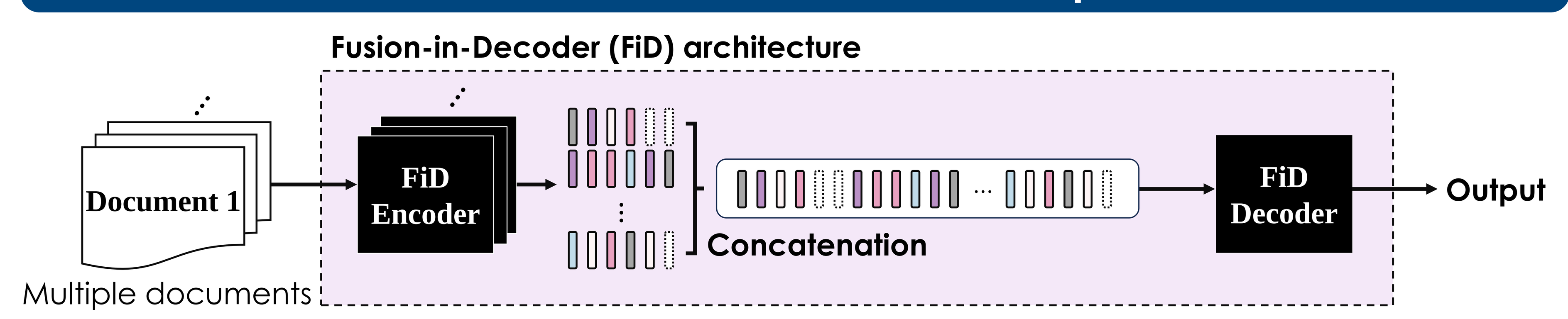


Challenge ② How to train a compressor effectively

- Incomplete pseudo-compressed prompts
- Non-differentiable tokens



Our Solution: Multi-document Reader as a Compressor

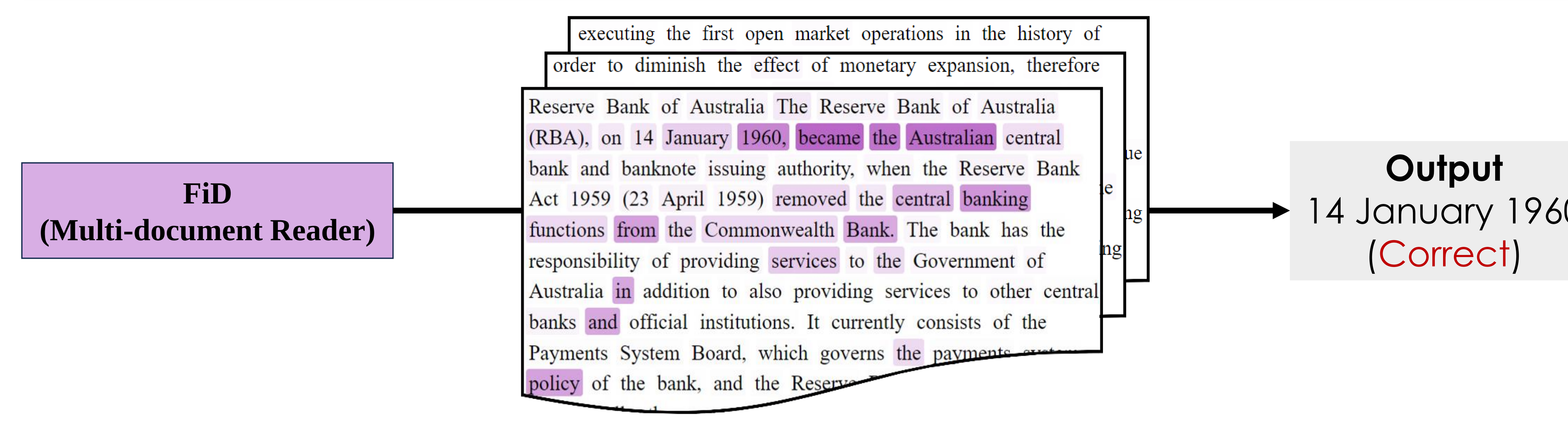


Multi-document reader aggregates evidence across multiple documents.

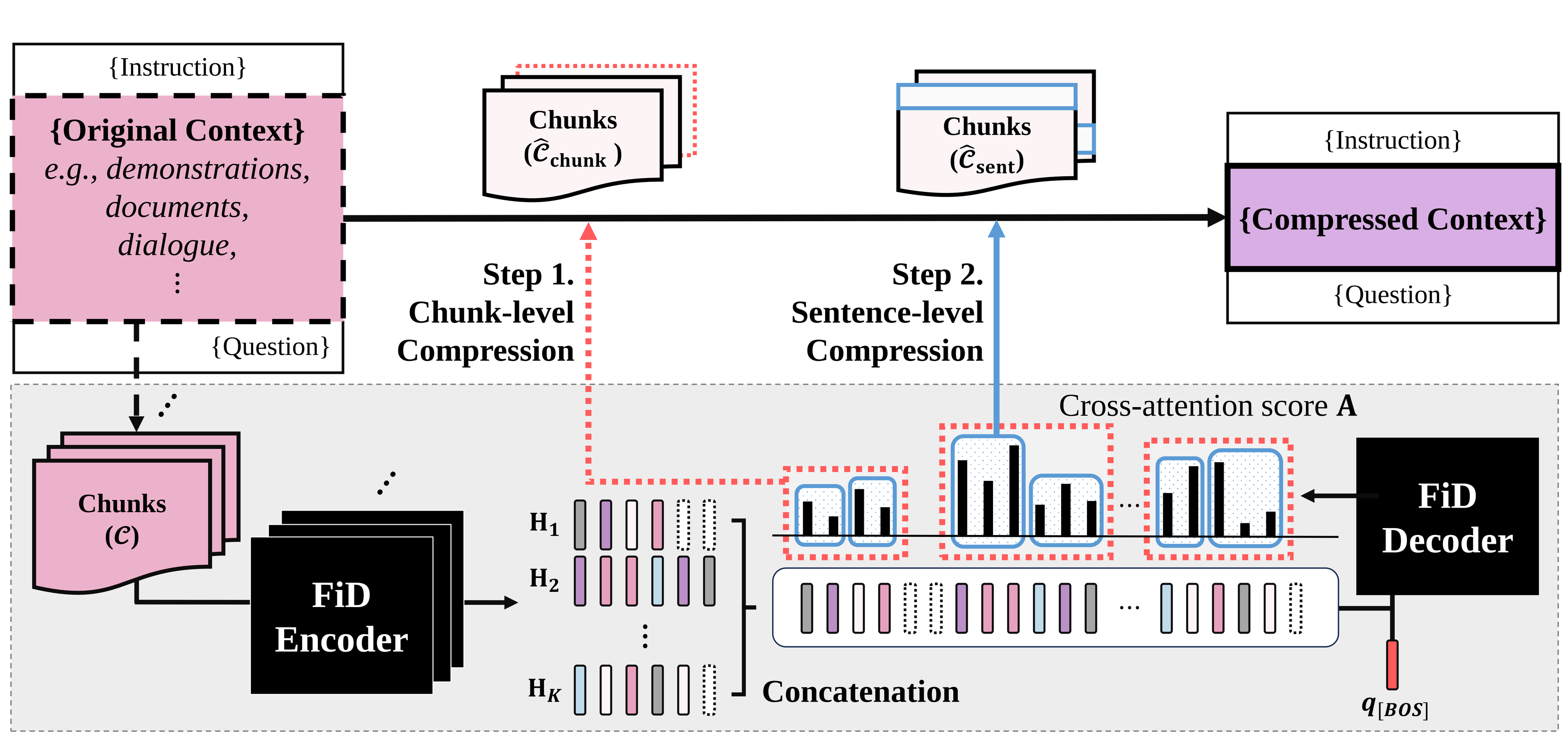
Addressing ①: **global context by using the concatenated KV matrix**

Addressing ②: **learning to capture by learning to answer the question**

- Avoiding the need for pseudo compressed prompts



Reading To Compressing (R2C)



Compression process

1) Encoding chunks

- Dividing long context into multiple chunks and put them into the FiD-encoder

2) Token-level importance

- Summing the cross-attention scores over all layers and heads in the FiD-decoder
- **Token-level importance of j-th token in i-th chunk** $t_{i,j} = \sum_{l=1}^{L}\sum_{h=1}^{H} \mathbf{A}_{i,j}^{(l,h)}$

3) Aggregating unit importance

- Token-level compression often **neglects the semantic integrity of the text.**

4) Hierarchical compression: R2C compresses the prompt hierarchically, given the multi-granularity unit importance, *i.e.*, chunk and sentence.

- Target number of tokens after compression T
- Hierarchical ratio between two levels of compression ρ

→ Number of tokens to compress $E_{comp} = |\mathrm{P_C}| - T$, where $|\mathrm{P_C}|$ is the original length.

✓ **Chunk-level compression**: $E_{chunk} = \rho \cdot E_{comp}$

✓ **Sentence-level compression**: $E_{sent} = (1-\rho) \cdot E_{comp}$

Experimental Results

In-domain Evaluation

- **Learning to answer the question** contributes to capture importance.
- A benefit of **using cross-attention scores**

Target LLM	Compression	NQ test (Span EM)	# tokens
FiD	-	50.5	-
GPT-3.5	Original	66.7	3,018
	BM25	49.4	534
	DPR	63.0	501
	Selective-Context	44.4	501
	LLMLingua-2	52.1	510
	LongLLMLingua	55.6	489
	RECOMP	63.0	500
	R2C (ours)	**66.5**	482
LLaMA2-7B	Original	52.8	3,018
	BM25	41.8	534
	DPR	54.3	501
	Selective-Context	38.1	501
	LLMLingua-2	42.5	510
	LongLLMLingua	49.0	489
	RECOMP	53.7	500
	R2C (ours)	**59.7**	482

Out-of-domain Evaluation

- **QA task proves to be an effective alternative** for training compressor.
- BM25 (Chunk-level) in FewShot task suggests the need to use varying granularity in compression.

Target LLM	Compression	SingleDoc	MultiDoc	Summ.	FewShot	Code	Avg.	# tokens
GPT-3.5	Original	43.2	46.1	25.2	69.2	64.4	49.6	9,881
	BM25	34.9	41.0	23.3	**68.1**	49.6	43.4	1,949
	Selective-Context	30.4	31.4	20.9	66.0	55.0	40.7	1,830
	LLMLingua	29.8	35.4	22.1	52.4	45.0	36.9	2,009
	LLMLingua-2	36.2	40.9	23.2	61.5	47.6	41.9	2,023
	LongLLMLingua	37.0	44.9	22.0	65.1	49.4	43.7	1,743
	RECOMP	40.1	48.1	-	-	-	-	-
	R2C (ours)	**43.5**	**48.7**	**24.9**	66.9	**57.6**	**48.3**	1,976

Compression Efficiency

R2C **dramatically improves compression efficiency**, while enhancing the accuracy.

Model	Backbone
LLMLingua	LLaMA2-7B
LongLLMLingua	LLaMA2-7B
LLMLingua-2	XLM-RoBERTa-large
R2C	T5-base