REFRAG: Rethinking RAG based Decoding

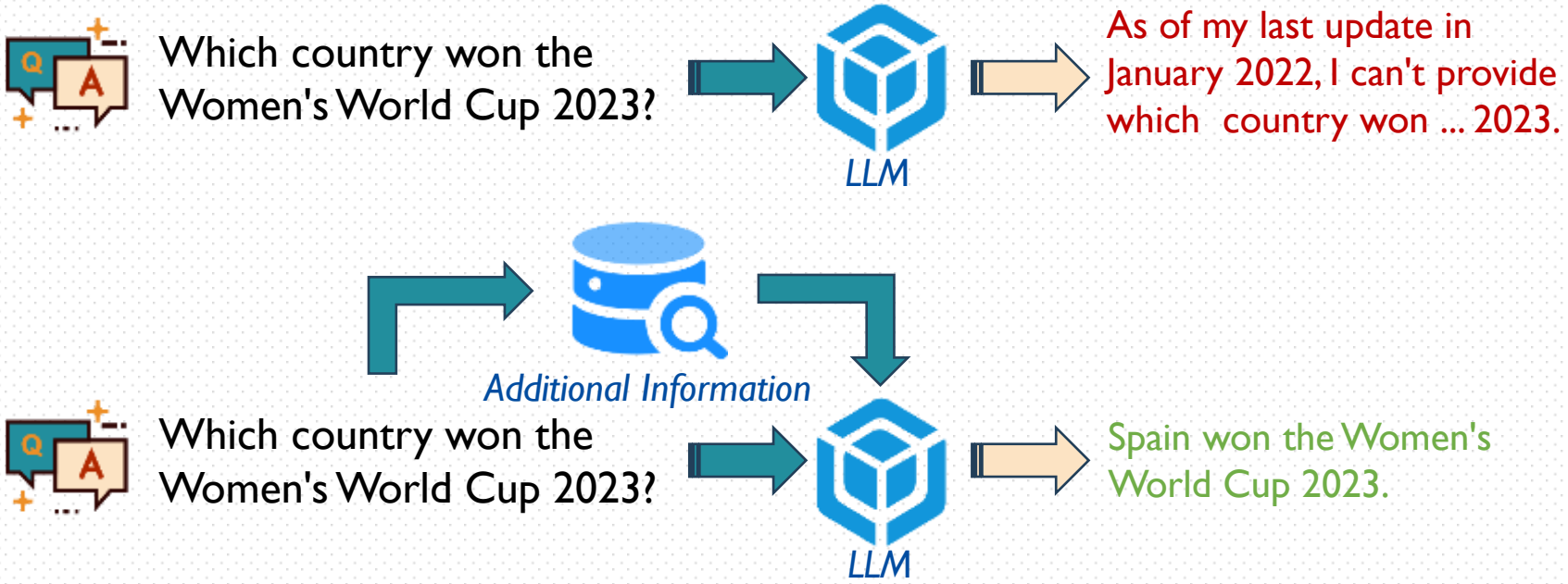
Xiaoqiang Lin[1,2,*], Aritra Ghosh[1], Bryan Kian Hsiang Low[2], Anshumali Shrivastava[1,3], Vijai Mohan[1]

[1]Meta Superintelligence Labs, [2]National University of Singapore, [3]Rice University

Bosen Yang @Reading Group 2026/01/13

RAG Framework

- RAG (Retrieval Augmented Generation)
  - Retrieve additional information with embedded query
  - Input the concatenation of retrieved context and query into LLM
- With RAG, LLMs generate more accurate answers



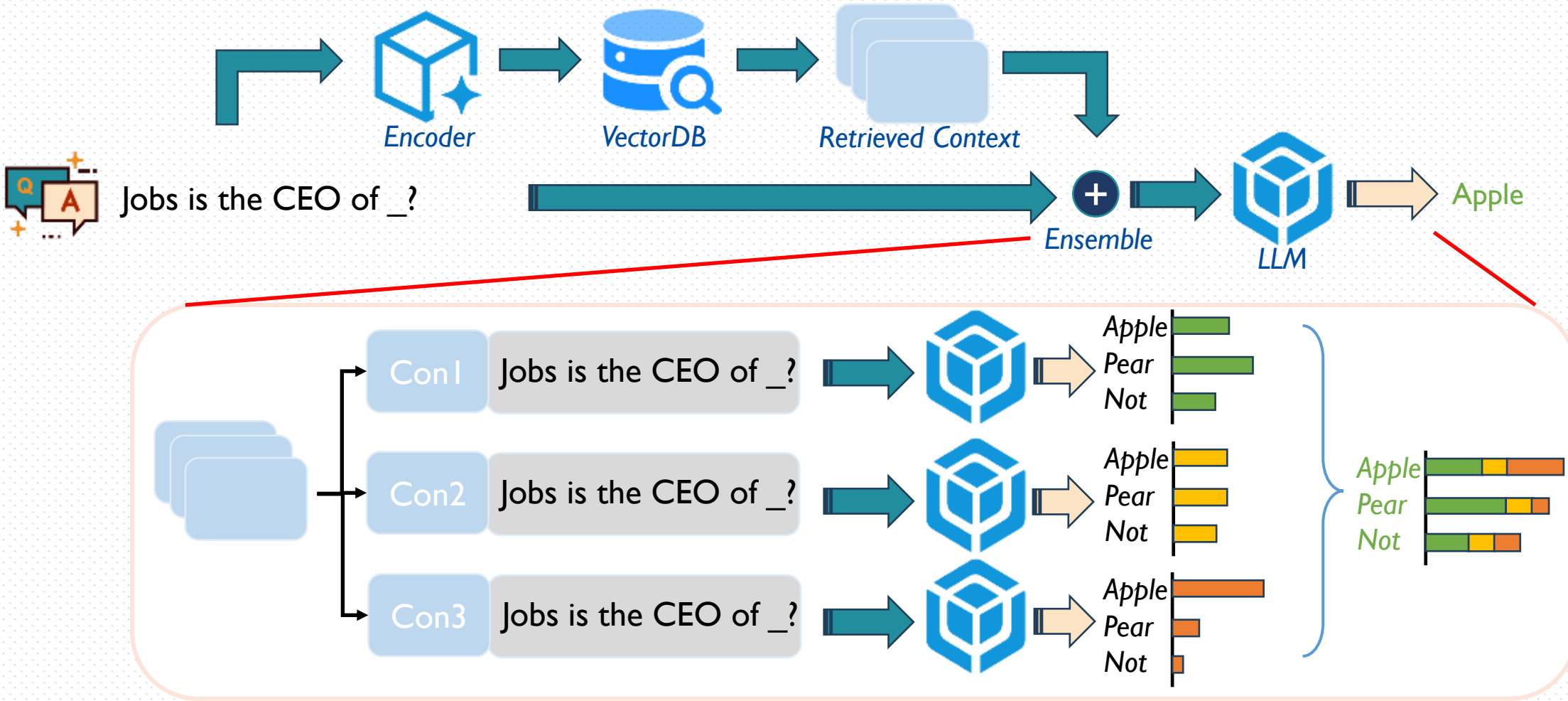
Optimization opportunities of RAG

- However, RAG suffers from longer context brought by retrieval
  - Longer inference latency
  - Additional memory consumption for KV cache
- RAG SOTA Works
  - REPLUG[1]
    - Alleviate context constraints by changing concatenation strategy
  - CEPE[2]
    - Improve RAG via parallel small-encoder processing and cross-attention integration

[1] REPLUG: Retrieval-Augmented Black-Box Language Models (NACCL'24)
[2] Long-Context Language Modeling with Parallel Context Encoding (ACL'24)

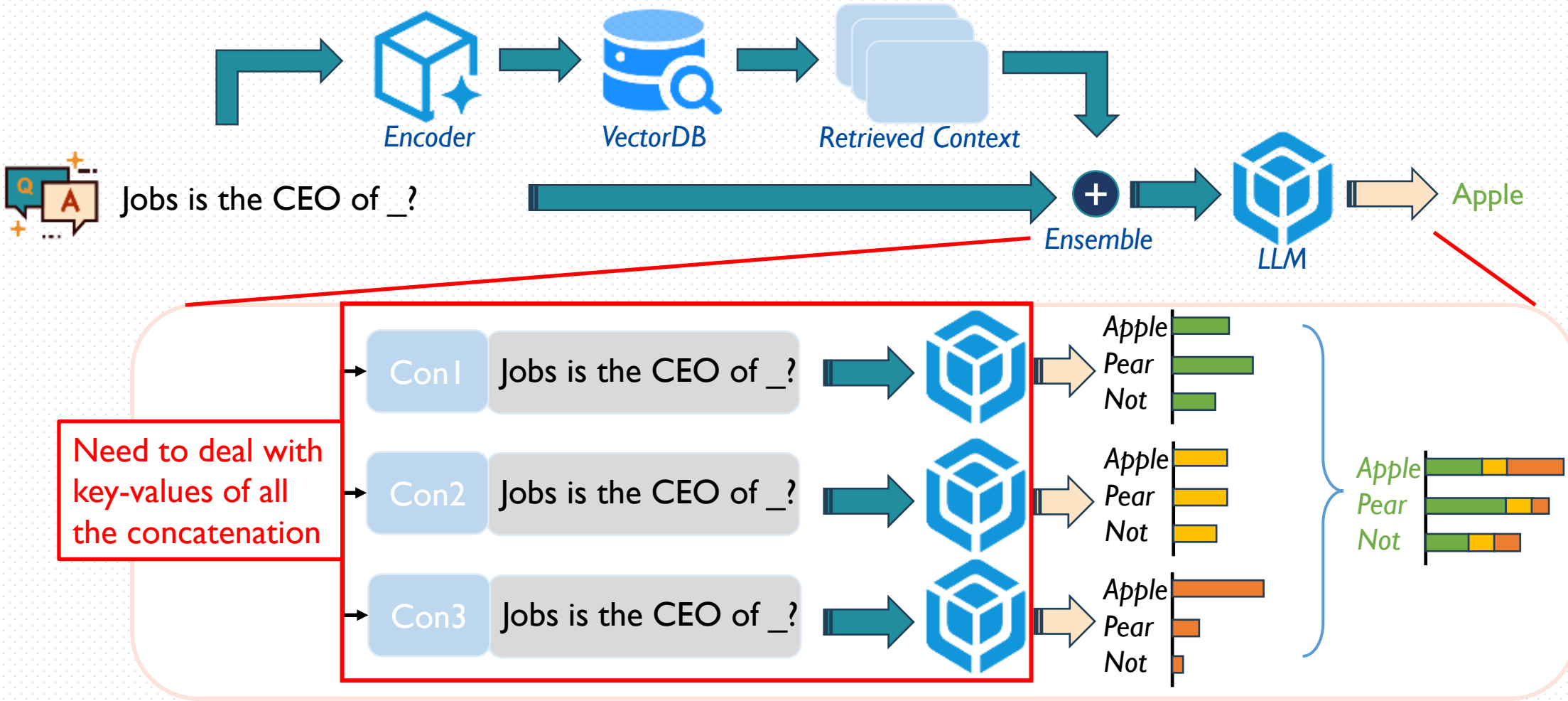
REPLUG Framework

- REPLUG ensembles output probabilities from different passes



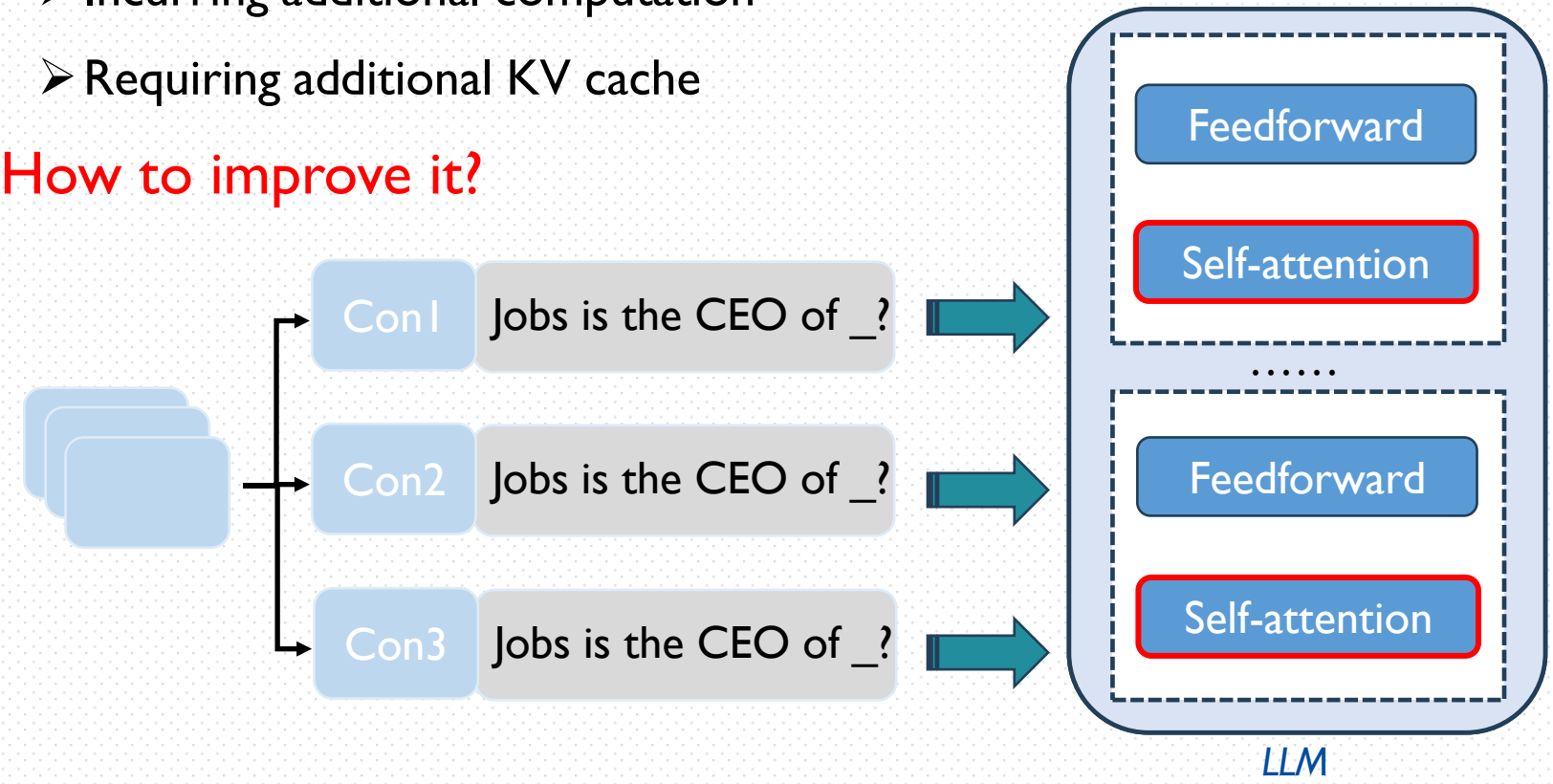
REPLUG Drawbacks

- REPLUG ensembles output probabilities from different passes



REPLUG is not good enough

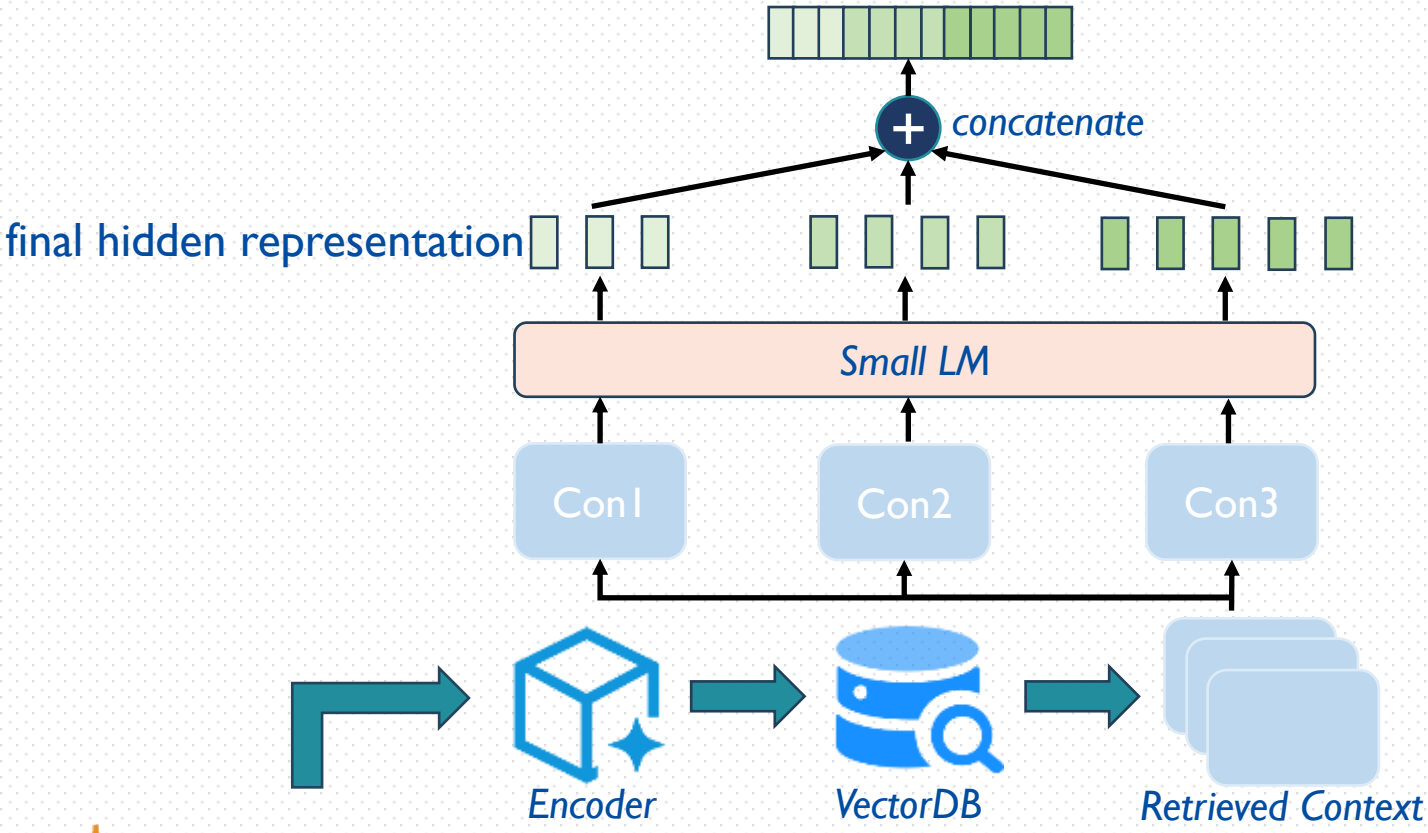
- REPLUG can transfer well to the long context setting
  - However, each chunk requires a forward pass of the main input[1]
    - Incurring additional computation
    - Requiring additional KV cache
  - How to improve it?



[1] Long-Context Language Modeling with Parallel Context Encoding (ACL'24)

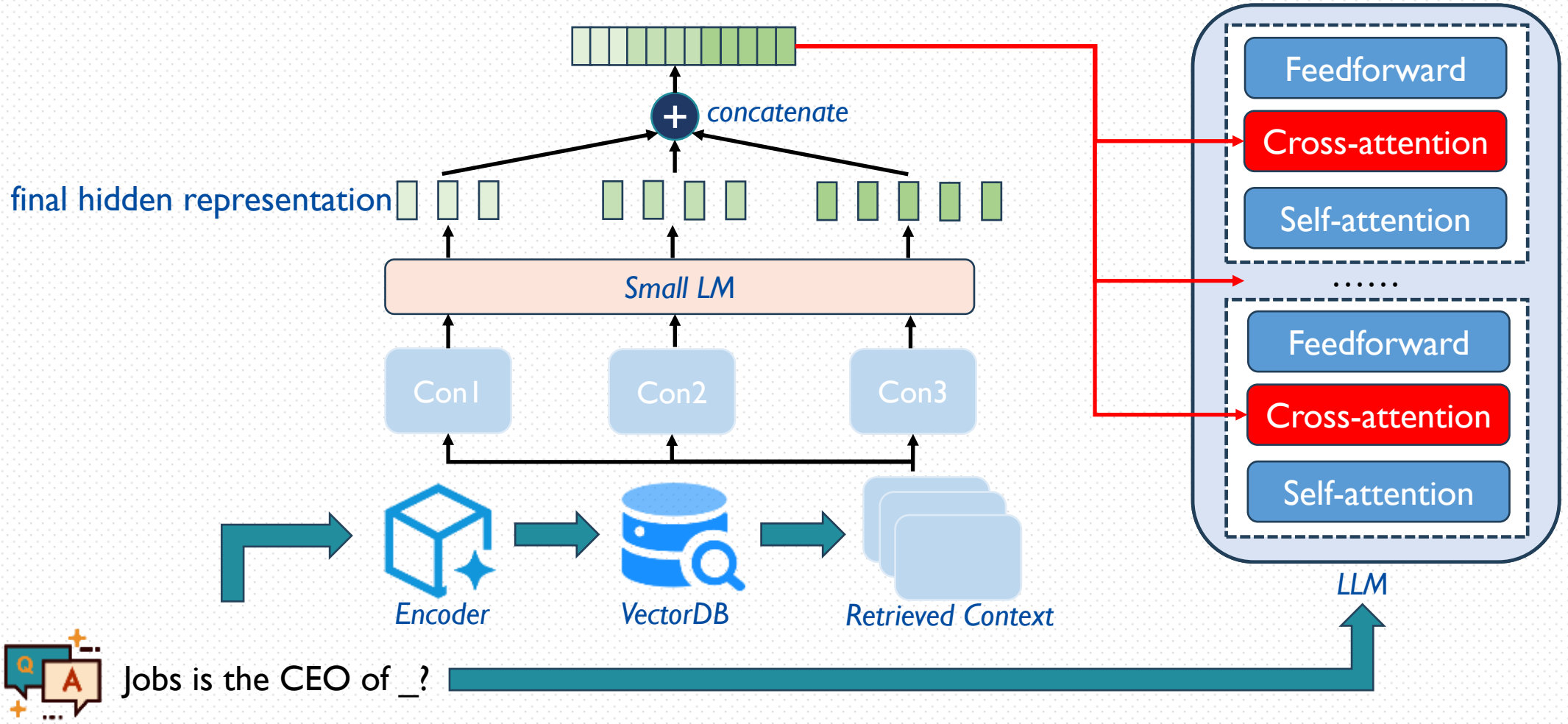
CEPE Framework

- CEPE extends long context via a parallel encoder and cross-attention



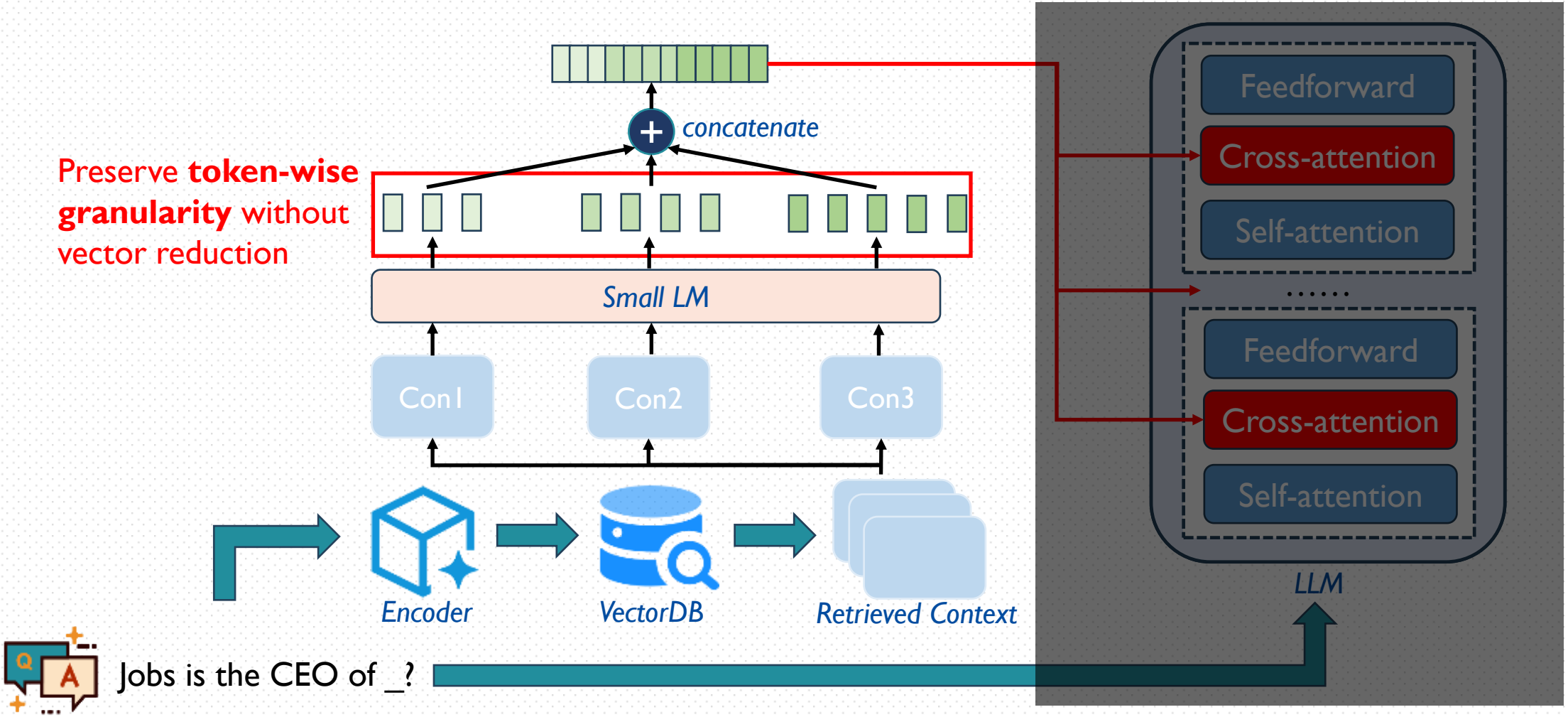
CEPE Framework

- CEPE extends long context via a parallel encoder and cross-attention



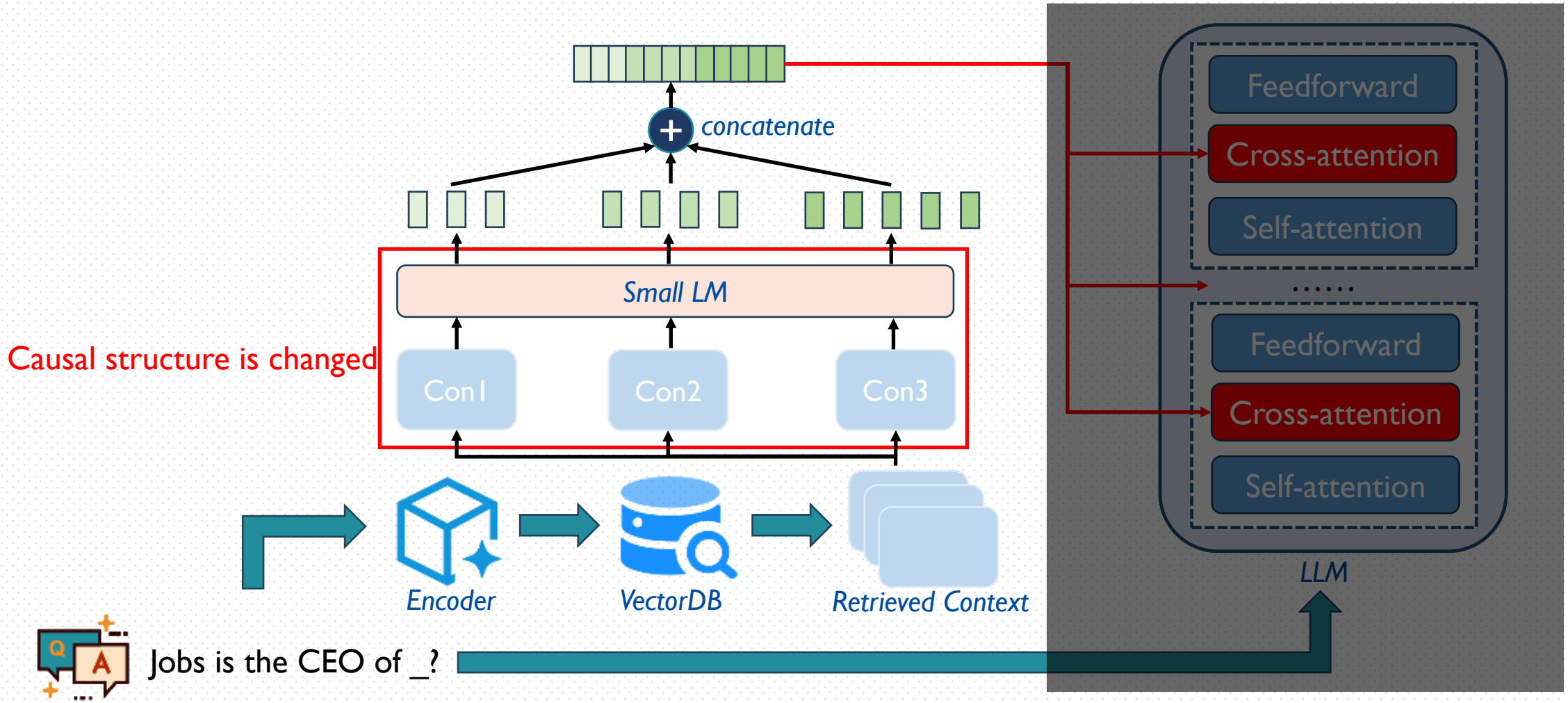
CEPE Drawbacks

- CEPE deals with chunks of context at fixed way



CEPE Drawbacks

- CEPE disrupts the causal structure



How to improve current works?

- REPLUG can transfer well to the long context setting
  - However, each chunk requires a forward pass of the main input[1]
    - Incurring additional computation
    - Requiring additional KV cache
- CEPE reduces both KV cache memory and attention computations
  - However, CEPE does not decrease numbers of embedded vectors
  - However, CEPE disrupts the causal structure of the context
- How to effectively deal with long context?
  - Compress each chunks of context with k tokens into *one* embedded vector

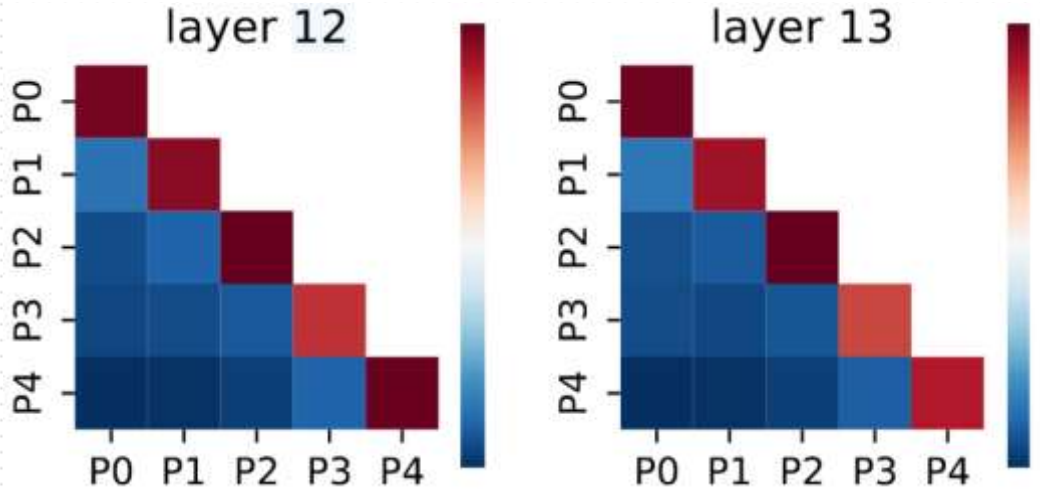
[1] Long-Context Language Modeling with Parallel Context Encoding (ACL'24)

The Case for Compression in RAG

- Inefficient token allocation
  - Many retrieved passages is uninformative and reused across multiple inferences
  - Allocating memory/computation for all the tokens is wasteful
  - Unusually structured and sparse attention
    - Most retrieved contexts during decoding are unrelated in RAG

Observations: Sparsity of retrieved passages

- Attention value visualization for different retrieved passages
  - Different layers for LLaMA-2-7B-Chat model
  - Pi denotes i-th retrieved passages



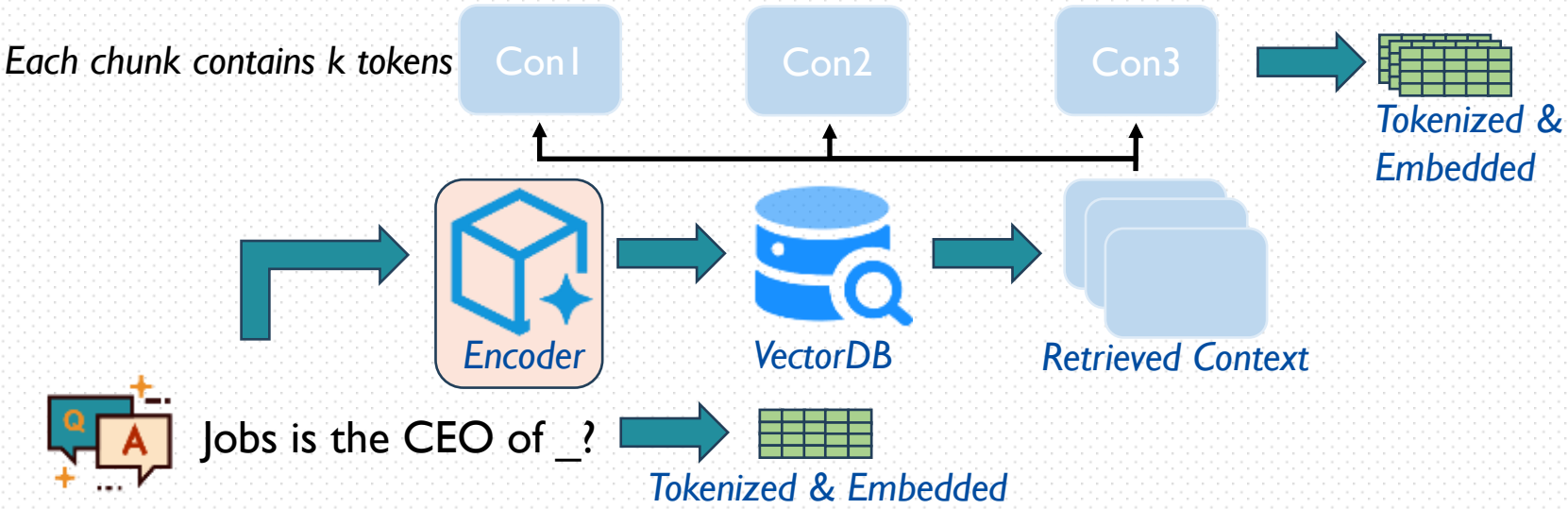
Attention value visualization for different layers

Challenge: How to effectively compress context

- How to decrease computation of the contexts?
- How to efficiently compress contexts?
- How to select useful contexts to keep accuracy?

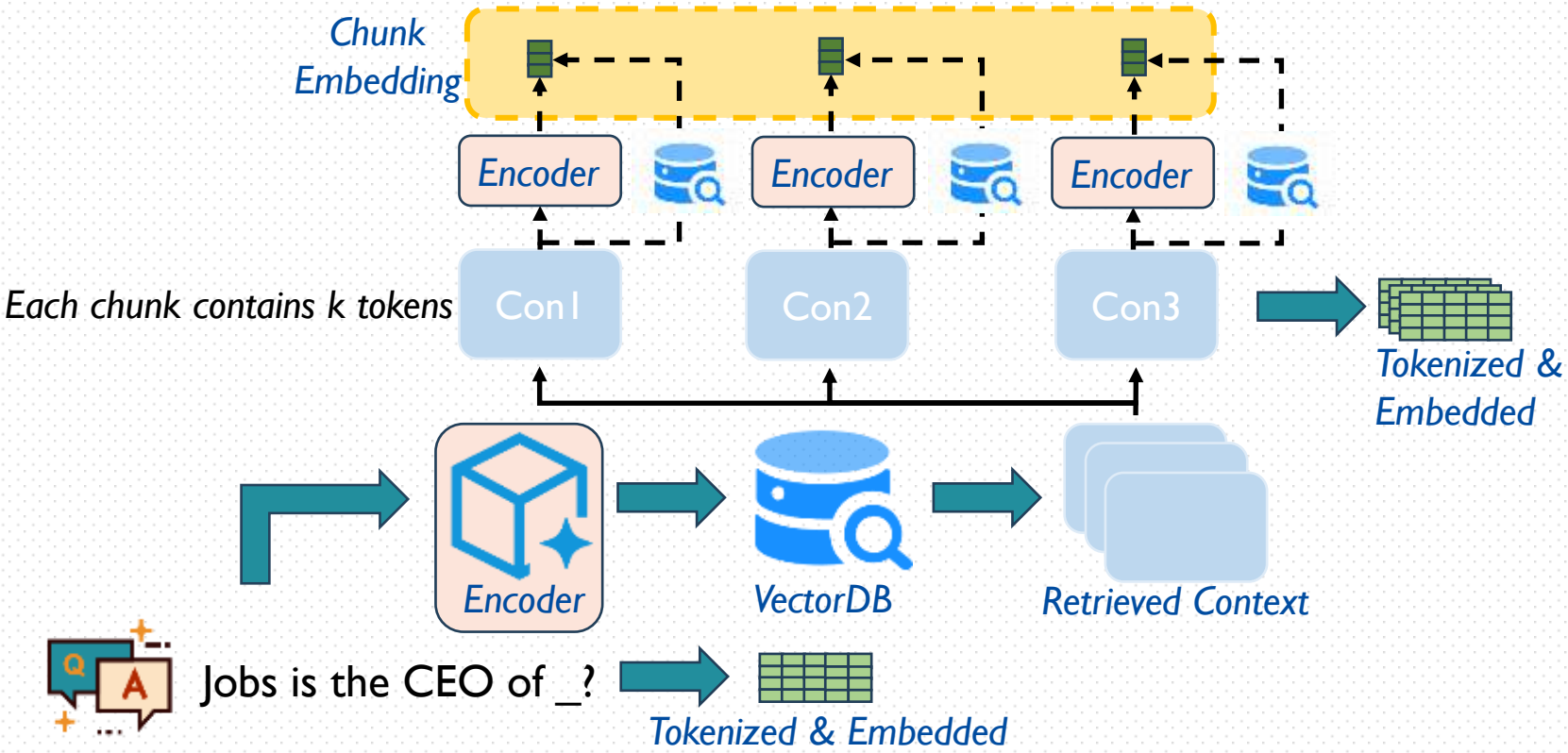
REFRAG Framework

- REFRAG splits context into chunks containing k tokens



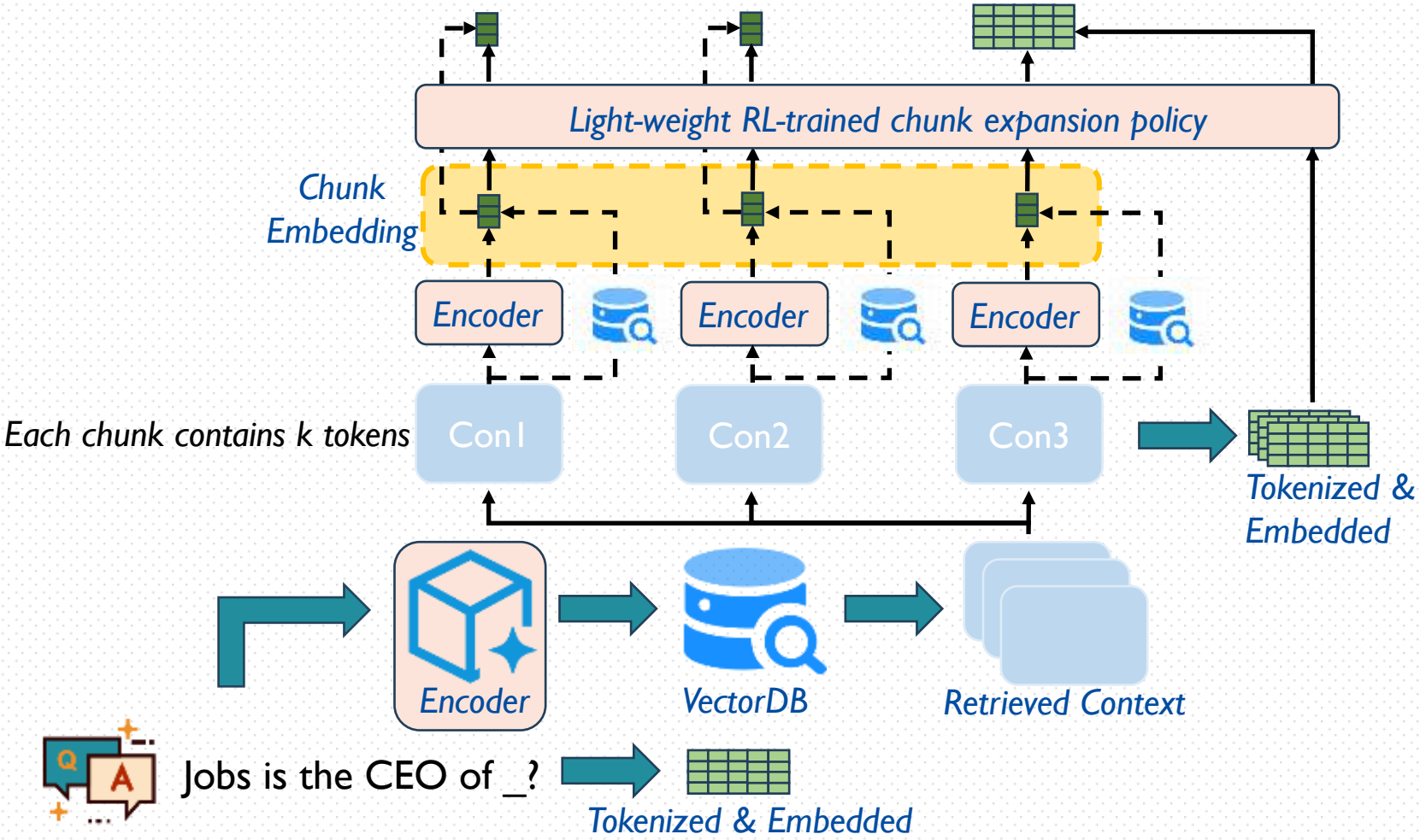
REFRAG Framework

- REFRAG obtains embeddings of each chunk as their compression
  - Embeddings can be cached



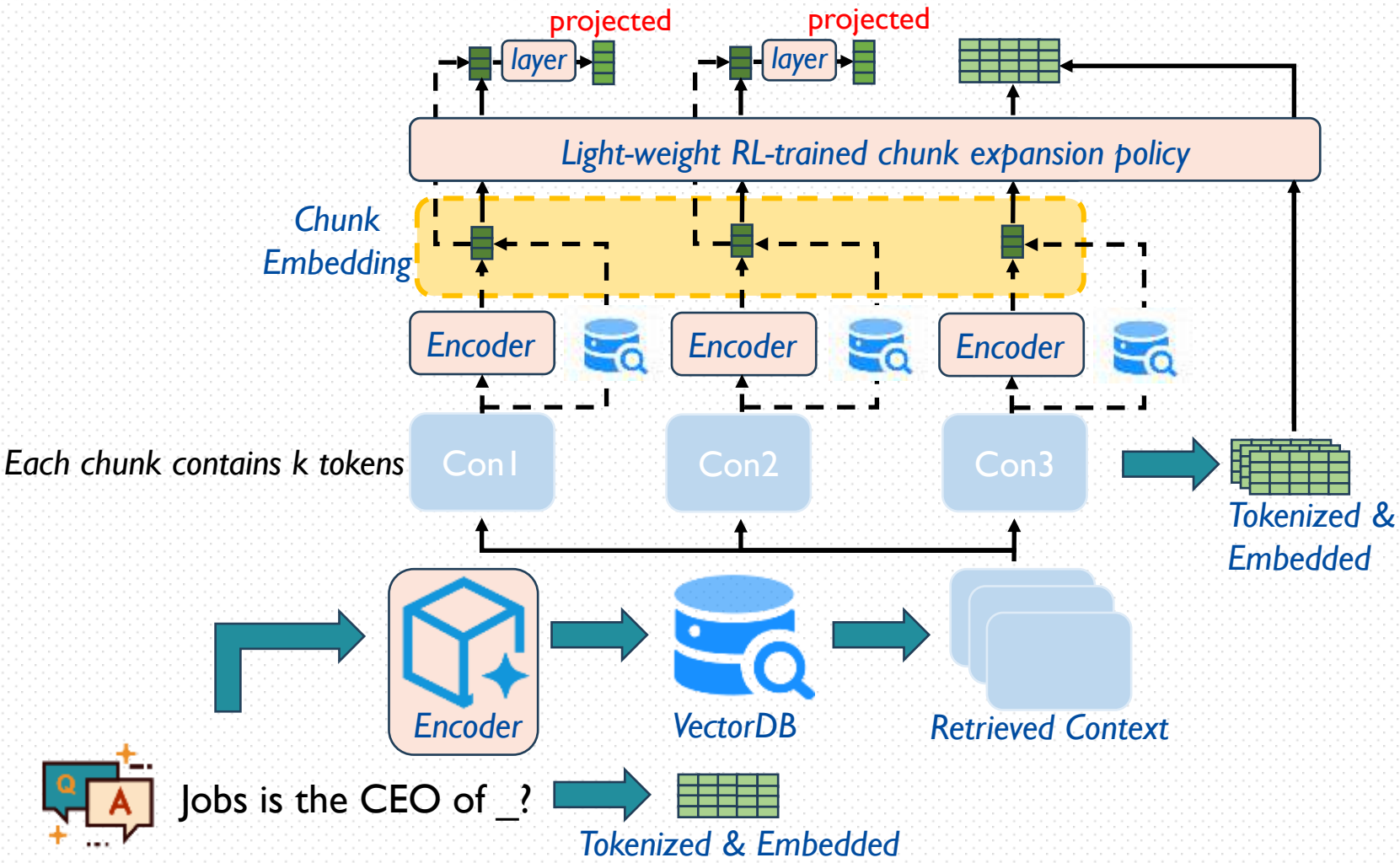
REFRAG Framework

- REFRAG expands useful context to be uncompressed



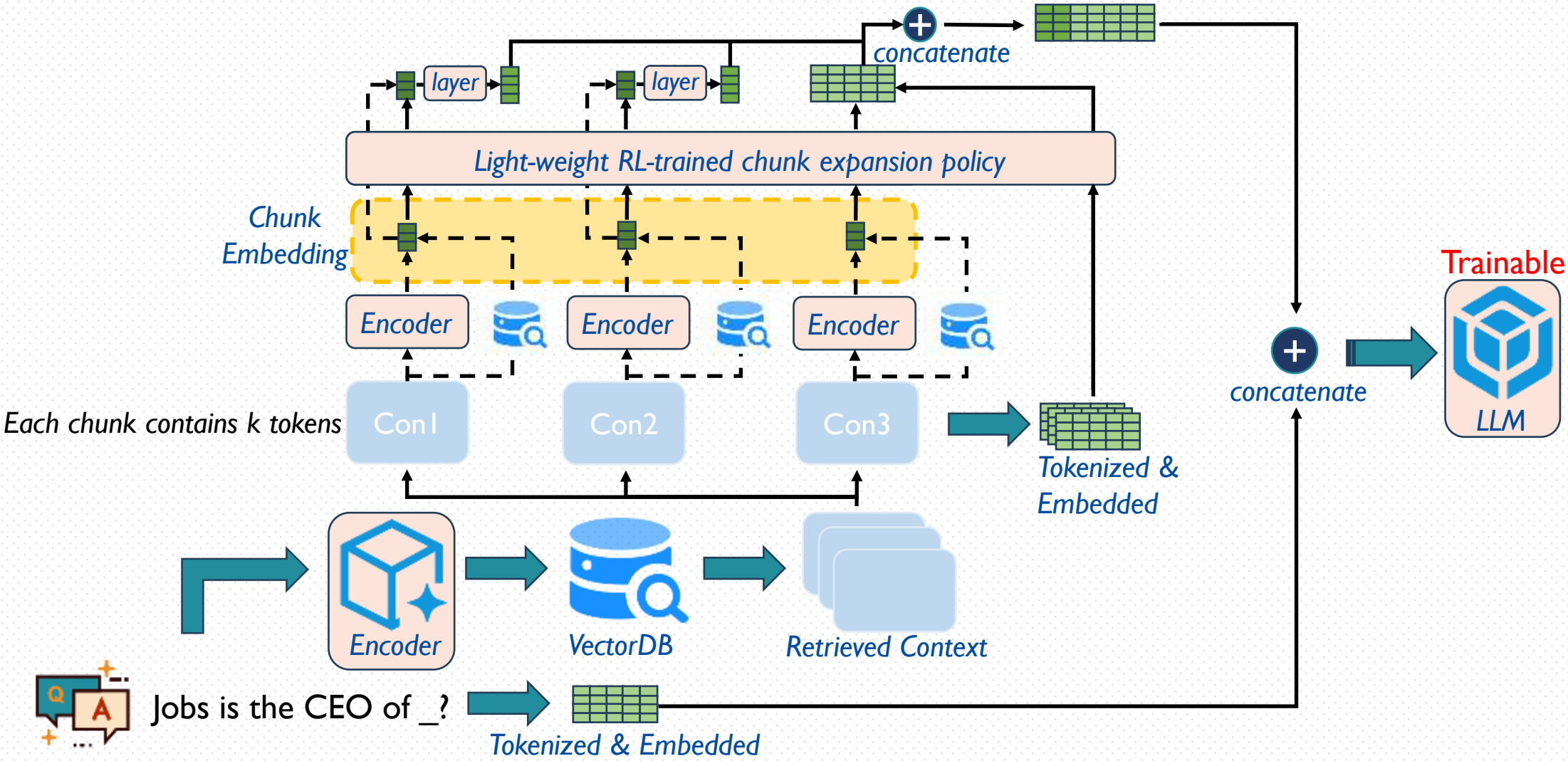
REFRAG Framework

- Projection Layer maps chunk embeddings into the decoder's token space



REFRAG Framework

- Decoder need to be fine-tuned



Methodology: Pre-training Encoder

- Reconstruction task
  - Freeze decoder
  - Train encoder and projection layer
    - Input s tokens into encoder and projection layer
    - Decoder **reconstructs** s tokens with the embeddings
    - Try to decrease Log-Perplexity of the output of decoder

$$\text{Log-Perplexity} = -\frac{1}{N}\sum_{i=1}^{N}\log P(x_i \mid \text{context}, x_1, \ldots, x_{i-1})$$

Methodology: Pre-training Decoder

- Continual pre-training decoder
  - Leverage trained encoder and projection layer
  - Training data contains mixture of compressed and uncompressed context
    - Each data point contains p fraction of compressed chunks of context
    - Adjust p to make decoder fit more difficult tasks
- Curriculum learning
  - Make training more effective
  - Gradually increase *Numbers/Difficulty* of training tasks

Methodology: Training RL policy

- Selective compression
  - Leverage trained encoder and decoder
  - Train RL policy to decide how to expand compressed chunks

		Arxiv			Book			PG19			ProofPile		
	Compression Rate	P512	P1024	P2048	P512	P1024	P2048	P512	P1024	P2048	P512	P1024	P2048 ↓
Context Length=2048													
REFRAG_8	8	1.124	1.091	1.062	1.905	1.868	1.844	1.996	1.956	**1.927**	0.997	0.952	0.916
$\text{REFRAG}_{16+\text{RL}}$	8.258	**1.118**	**1.090**	1.062	**1.878**	**1.856**	**1.840**	**1.978**	**1.952**	1.930	**0.992**	**0.951**	0.916
Context Length=4096													
REFRAG_8	8	1.098	1.065	1.042	1.895	1.860	1.837	1.989	1.950	1.922	0.965	0.923	0.894
$\text{REFRAG}_{16+\text{RL}}$	8.0157	**1.065**	**1.048**	**1.033**	**1.851**	**1.837**	**1.828**	**1.952**	**1.934**	**1.918**	**0.932**	**0.905**	**0.883**

Performance comparison with and w/o RL policy

Evaluation Setup

- REFRAG Model
  - Encoder
    - RoBERTa
  - Decoder
    - LLaMA-2
- Evaluation Situation
  - Normal Generation
  - RAG
  - Multi-Turn Conversation

Evaluation Setup

- Baselines
  - LLaMA-No Context – Perform Worst
  - LLaMA-Full Context/LLaMA-32K – Perform Best
  - LLaMA_K
  - REPLUG
  - CEPE

Evaluation: Normal Generation

- Fixed context with variable output lengths
  - Context s = 2048, Output o ∈ {512, 1024, 2048}
  - REFRAG performs best excluding LLaMA-Full Context/LLaMA-32K

	Arxiv			Book			PG19			ProofPile		
	P512	P1024	P2048	P512	P1024	P2048	P512	P1024	P2048	P512	P1024	P2048 ↓
LLaMA-Full Context	1.075	1.074	1.069	1.830	1.827	1.826	1.947	1.941	1.935	0.952	0.940	0.931
LLaMA-32K	1.086	1.084	1.076	1.887	1.883	1.880	1.988	1.982	1.975	0.961	0.948	0.937
LLaMA-No Context	1.526	1.371	1.254	2.101	1.995	1.927	2.211	2.102	2.030	1.437	1.256	1.127
LLaMA_{256}	1.267	1.221	1.171	**1.924**	1.897	1.874	2.031	2.003	1.978	1.156	1.094	1.038
REPLUG	1.526	1.371	1.254	2.101	1.995	1.927	2.211	2.102	2.030	1.437	1.256	1.127
CEPE	1.196	1.148	1.107	1.946	1.896	1.864	2.057	2.002	1.964	1.075	1.014	0.968
REFRAG_8	**1.124**	**1.091**	**1.062**	**1.905**	**1.868**	**1.844**	**1.996**	**1.956**	**1.927**	**0.997**	**0.952**	**0.916**
REFRAG_{16}	**1.157**	**1.114**	**1.076**	1.925	**1.882**	**1.853**	**2.016**	**1.971**	**1.938**	**1.034**	**0.976**	**0.931**
REFRAG_{32}	1.215	1.154	**1.103**	1.946	**1.896**	**1.862**	2.039	**1.987**	**1.949**	1.097	1.020	**0.961**

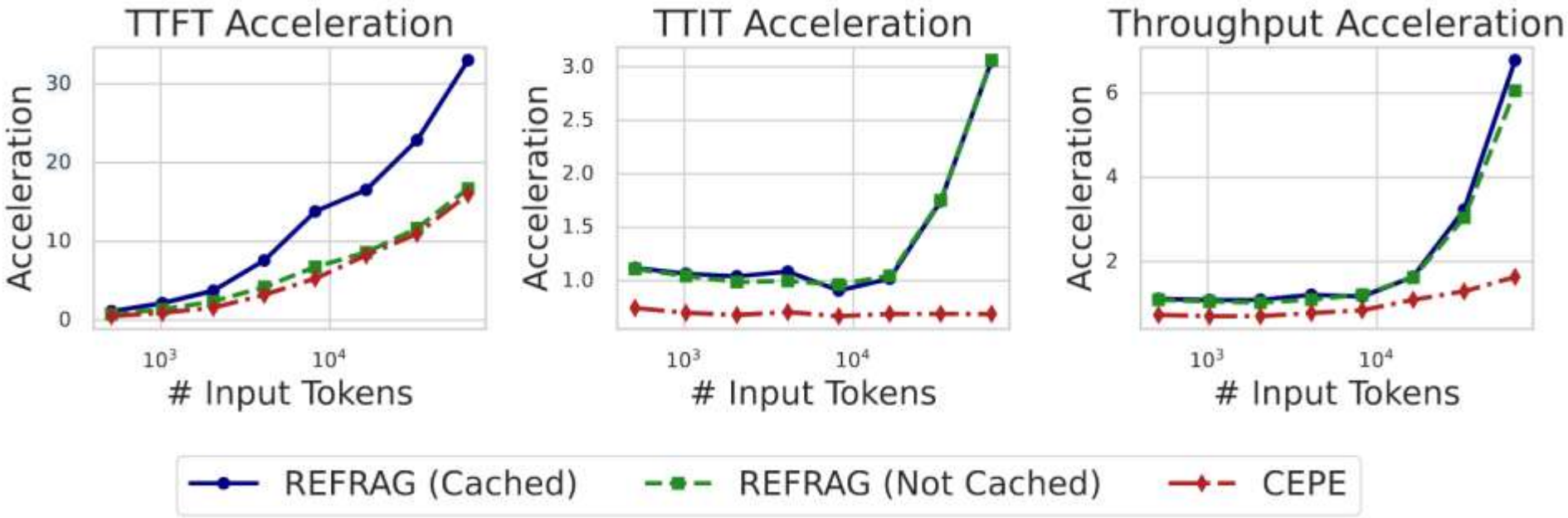
Evaluation: Normal Generation

- Variable context with fixed output lengths
  - Context s ∈ {4096, 8192, 16384}, Output o = 2048
  - REFRAG enables extrapolation of context window

	Context Length =4096				Context Length=8192				Context Length=16384			
	Arxiv	Book	PG19	ProofPile	Arxiv	Book	PG19	ProofPile	Arxiv	Book	PG19	ProofPile ↓
LLaMA-Full Context	6.751	6.956	6.829	6.701	9.675	9.069	8.963	9.401	9.043	8.913	8.848	8.989
LLaMA-32K	1.037	1.862	1.960	0.898	0.965	1.867	1.947	0.834	0.865	1.840	1.943	0.770
LLaMA-No Context	1.253	1.925	2.030	1.126	1.226	1.949	2.032	1.110	1.174	1.939	2.041	1.081
REPLUG	1.253	1.925	2.030	1.126	1.226	1.949	2.032	1.110	1.174	1.939	2.041	1.081
CEPE	1.085	1.856	1.959	0.945	1.032	1.878	1.958	0.904	0.960	1.864	1.966	0.863
REFRAG_8	**1.042**	**1.837**	**1.922**	**0.894**	**0.983**	**1.839**	**1.922**	**0.858**	0.977	**1.840**	**1.939**	0.891
REFRAG_{16}	**1.058**	**1.847**	**1.934**	**0.910**	**0.994**	**1.845**	**1.932**	**0.871**	**0.942**	**1.840**	**1.945**	**0.850**
REFRAG_{32}	1.088	1.857	**1.946**	**0.944**	**1.032**	**1.860**	**1.945**	0.912	0.969	**1.852**	**1.955**	0.880

Evaluation: Normal Generation

- Variable context with fixed output lengths
  - Acceleration ranges from *k(short)* to k^2*(long)*
  - Without cache, encoding costs



Empirical verification of inference acceleration of REFRAG with k = 16

Evaluation: RAG

- REFRAG outperforms other models
  - REFRAG performs well under the same/less latency
  - REFRAG enables extracting more useful information

Multi-Choice	MMLU	CommonsenseQA	MathQA	ECQA	HellaSwag	SIQA	PIQA	Winogrande ↑	(1/ # tokens)
Short context with the same latency									
LLAMA$_{FT}$ + 1 context	**50.23**	85.05	99.50	84.77	41.80	68.12	67.36	55.64	1×
REFRAG$_8$ + 8 passages	50.29	**92.27**	**99.66**	**94.70**	**45.23**	**68.94**	**71.38**	**57.70**	1×
REFRAG$_{16}$ + 8 passages	49.84	**89.18**	**99.66**	**98.01**	39.33	**68.42**	**70.29**	**56.67**	2×
REFRAG$_{32}$ + 8 passages	49.51	**91.75**	**99.50**	**97.35**	**42.86**	**68.17**	**68.34**	**56.75**	4×
Long context									
LLAMA$_{FT}$ + 10 passages	48.66	82.99	68.46	84.11	41.77	67.45	68.01	53.91	1×
CEPED +80 passages	26.26	26.29	23.66	24.50	24.95	32.86	48.53	44.51	
REPLUG +80 passages	-	78.35	-	76.16	-	65.51	-	-	
LLAMA-32K +80 passages	22.21	16.49	19.80	16.56	23.76	24.16	34.17	48.86	
REFRAG$_8$ +80 passages	**50.42**	**92.27**	**99.66**	**97.35**	**44.61**	**68.22**	**69.37**	**57.54**	1×
REFRAG$_{16}$ +80 passages	**50.88**	**89.69**	**99.66**	**96.69**	38.50	**68.47**	**70.89**	**56.99**	2×
REFRAG$_{32}$ +80 passages	**49.77**	**90.72**	**99.50**	**98.01**	**43.37**	**68.47**	**69.04**	**56.99**	4×

- means the corresponding model has out-of-memory error.

Evaluation: Multi-Turn Conversation

- Retrieve N passages for each conversation turn
  - LLaMA necessitates truncating portions of the long conversational history
  - REFRAG maintains robust performance owing to its compression way

	# Turns (≥)	ORConvQA	QReCC	TopiOCQA ↑
	# Passages = 5			
LLaMA_{FT}	2	20.73	**18.72**	26.98
REFRAG_{8}	2	**21.17**	17.73	28.04
REFRAG_{16}	2	20.19	17.30	27.89
REFRAG_{32}	2	19.70	17.35	**28.67**
LLaMA_{FT}	4	20.33	**16.42**	23.50
REFRAG_{8}	4	22.78	15.61	26.93
REFRAG_{16}	4	**21.94**	15.27	**27.03**
REFRAG_{32}	4	21.68	15.45	26.45
LLaMA_{FT}	6	20.76	**11.92**	23.10
REFRAG_{8}	6	**23.11**	10.88	25.37
REFRAG_{16}	6	21.69	10.75	**26.17**
REFRAG_{32}	6	21.19	10.69	25.51

	# Turns (≥)	ORConvQA	QReCC	TopiOCQA ↑
	# Passages = 10			
LLaMA_{FT}	2	16.52	17.31	23.02
REFRAG_{8}	2	**21.15**	**17.92**	27.97
REFRAG_{16}	2	20.79	17.37	**28.45**
REFRAG_{32}	2	19.67	17.16	28.31
LLaMA_{FT}	4	16.90	14.69	20.23
REFRAG_{8}	4	**22.63**	**15.68**	25.95
REFRAG_{16}	4	21.84	15.21	**26.12**
REFRAG_{32}	4	21.75	15.33	25.77
LLaMA_{FT}	6	14.44	10.72	19.52
REFRAG_{8}	6	20.59	**11.00**	25.16
REFRAG_{16}	6	21.05	10.50	24.96
REFRAG_{32}	6	**21.67**	10.79	**25.23**

Conclusion

- Highlights
  - Reuse precomputable results of encoder
  - Preserve the autoregressive nature of the decoder
  - Compress chunks of context at arbitrary positions
  - Select useful context to be uncompressed to keep accuracy

- Potential problems
  - Length of context in experiments is not long enough