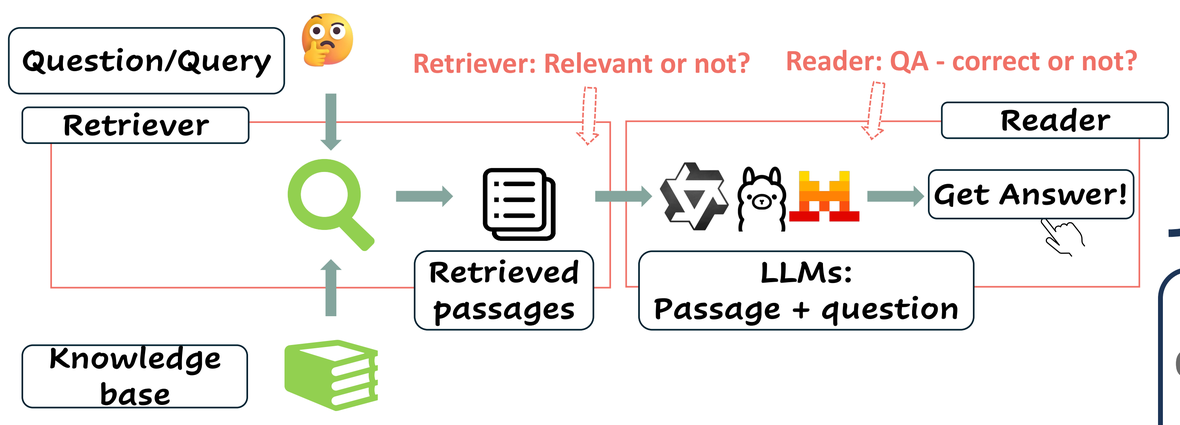
Beyond Prompting: An Efficient Embedding Framework for Open-Domain Question Answering

Zhanghao Hu[1], Hanqi Yan[1], Qinglin Zhu[1], Zhenyi Shen[1], Yulan He[1,2], Lin Gui[1]
[1]King's College London, [2]The Alan Turing Institute

ACL 2025 VIENNA

Motivation: Limitation of Retriever- Reader Framework in ODQA System

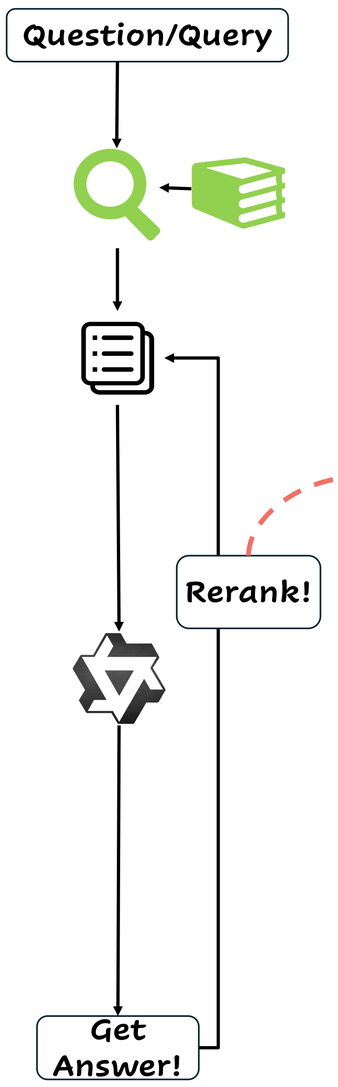


Question 1: Retriever----How to prioritize passages that contain ground-truth answers with efficient rerank retriever?

Question 2: Reader---- How to reduce multi-turn prompts of LLMs to predict final answer?

Retriever: Self-Refinement Driven Reranking

Self-refinement: focus on the initial answer and rerank the relevant documents



Example

Q: Which of the actors starred in "Sollathaan Ninaikkiren" was born in 7 November 1954?

Candidate answer:
A. Sollathaan Ninaikkiren (wrong), Kamal Haasan (correct)

Principle of reranking: Elevating relevant passages and suppressing irrelevant ones to recalibrate candidate rankings across the refinement query embedding by a linear projection.

Original order

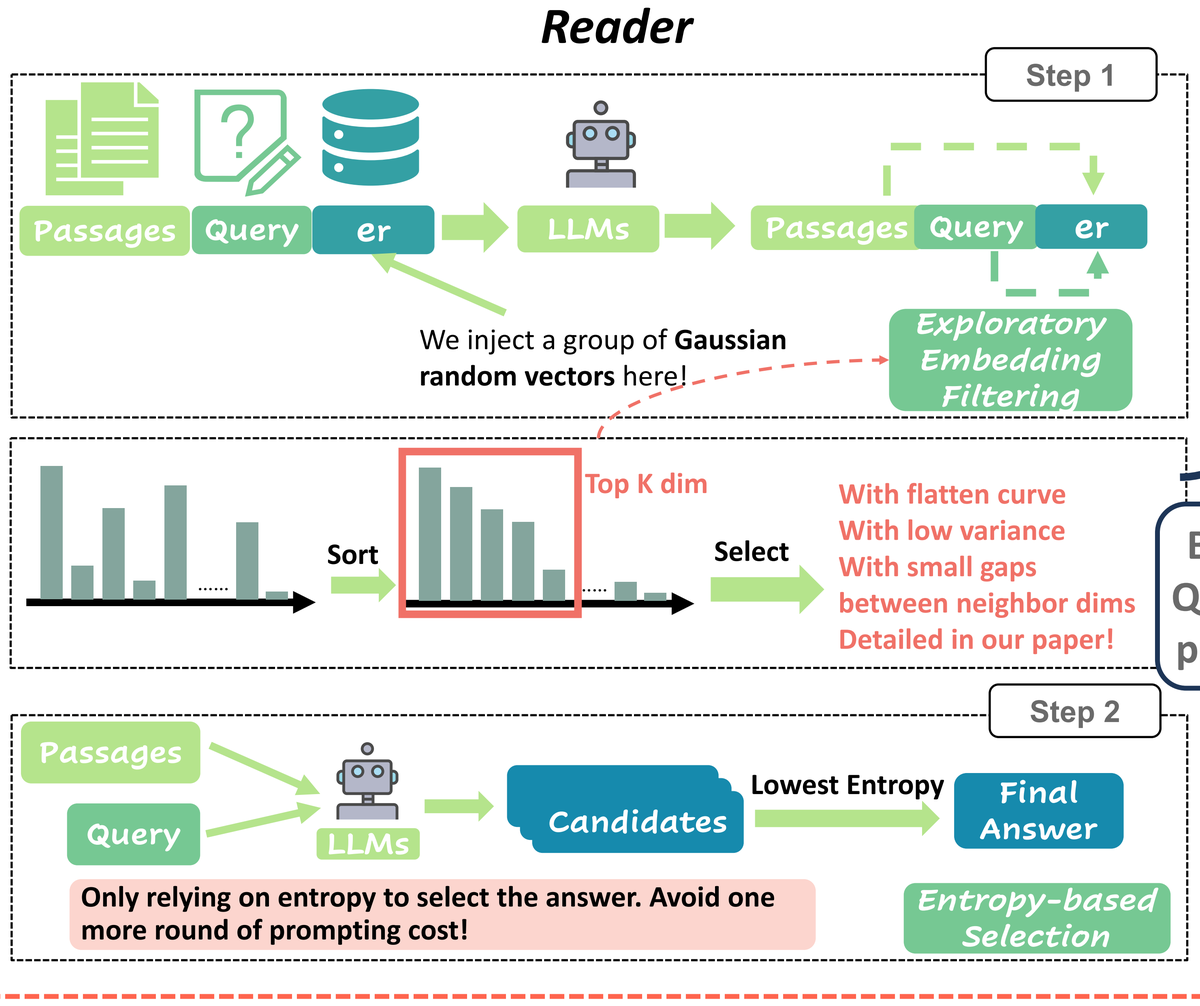
1. ... Kamal Haasan played a guest role in the film.....
2. ... Sollathaan Ninaikkiren is a 1973 Tamil language .....
3. ... Johanna Raunio (born 5 November 1954 in Turku) is a Finnish actress.....
4. ... Sundara Swapnagalu is a 1986 Kannada melodrama ...
5. ... In the song, the main character Balakrishna, played by Kamal Haasan.....

Rerank

Updated order

1. ..... Kamal Haasan played a guest role in the film.....
2. ..... In the song, the main character Balakrishna, played by Kamal Haasan.....
3. ... Sollathaan Ninaikkiren is a 1973 Tamil language .....
4. ... Johanna Raunio (born 5 November 1954 in Turku) is a Finnish actress.....
5. ... Sundara Swapnagalu is a 1986 Kannada melodrama film .....

Reader: Enhancing Generation via Exploratory Embedding



Better QA, less cost.

Better rerank, better QA.

Better QA, less prompt

Experiment Results

(1). EmbQA Outperforms Prompt-Level Methods & Efficiency.

Method/Dataset	HotpotQA		2Wiki		NQ		WebQ		Average	
	EM	F1	EM	F1	EM	F1	EM	F1	EM	F1
No Retrieval	20.8	29.1	12.2	16.2	20.6	26.6	17.2	25.8	17.7	24.4
Retrieval Only	25.4	37.2	16.6	21.1	26.0	32.8	22.2	31.2	22.6	30.6
Chain-of-Thought	27.0	39.8	15.4	21.8	27.2	33.5	28.8	37.8	24.6	33.2
Self-Verification	32.8	49.5	21.0	23.5	28.0	37.7	27.2	40.2	27.4	38.0
SuRe	38.8	53.5	23.8	31.0	36.6	47.9	34.4	48.5	33.4	45.3
RPG	37.9	49.2	24.6	33.8	36.6	50.5	34.2	47.3	33.3	45.2
KnowTrace	38.8	48.9	24.7	33.5	33.7	43.1	32.2	44.7	32.4	42.6
EmbQA (Ours)	**42.0**	**55.8**	**27.4**	**36.6**	**42.2**	**54.4**	**38.2**	**52.1**	**37.5**	**49.7**

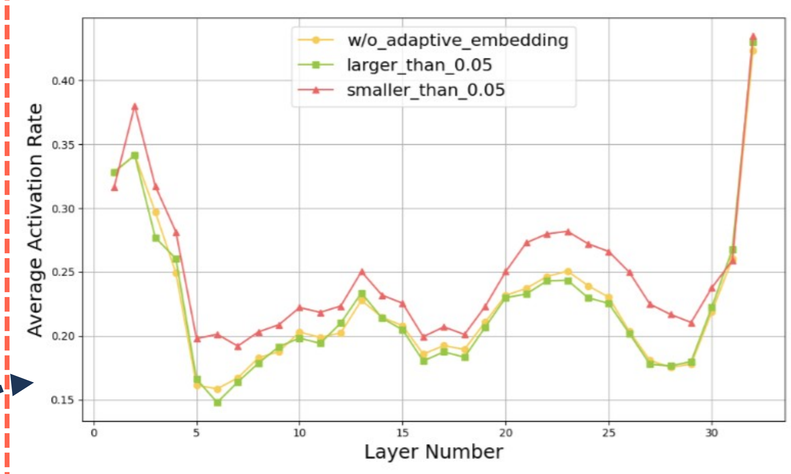
Dataset	Method	Time/query (min) ↓	Tokens /query ↓
HotpotQA	SuRe	1.56	3.51k
	EmbQA (ours)	**0.53**	**0.99k**
2Wiki	SuRe	1.57	3.43k
	EmbQA (ours)	**0.54**	**1.20k**
NQ	SuRe	1.43	4.39k
	EmbQA (ours)	**0.54**	**0.84k**
WebQ	SuRe	1.58	3.91k
	EmbQA (ours)	**0.56**	**1.31k**

(2). Why Prompt-Level Re-rank Framework Fail in Existing ODQA Framework?

Retriever& Rerank Framework	HotpotQA		2Wiki		NQ		WebQ	
	Avg. GT @Top-10	Time /Query(s)	Avg. GT @Top-10	Time /Query(s)	Avg. GT @Top-10	Time /Query(s)	Avg. GT @Top-10	Time /Query(s)
BM25	1.16	–	0.81	–	1.50	–	1.88	–
+Prompt Level	1.06	12.52	1.09	12.62	1.62	12.65	2.70	12.69
+Embedding Level (Ours)	**1.42**	**1.33**	**1.21**	**1.54**	**2.57**	**1.90**	**4.18**	**2.31**
DPR	0.28	–	0.30	–	1.79	–	3.04	–
+Prompt Level	0.64	13.23	0.34	12.52	1.75	12.66	3.68	12.63
+Embedding Level (Ours)	**1.01**	**2.42**	**1.13**	**1.27**	**2.41**	**2.00**	**4.25**	**2.22**
Contriever	1.47	–	0.99	–	1.98	–	2.87	–
+Prompt Level	1.37	12.54	1.36	12.95	2.01	13.16	3.02	13.05
+Embedding Level (Ours)	**1.87**	**1.12**	**1.49**	**2.93**	**2.55**	**2.12**	**4.31**	**2.69**

Retriever & Rerank Module	HotpotQA		2Wiki		NQ		WebQ	
	EM	F1	EM	F1	EM	F1	EM	F1
BM25	25.4	37.2	16.6	21.1	26.0	32.8	22.2	31.2
+ Prompt-Level	39.6	52.4	19.8	28.7	39.8	51.8	36.6	50.1
+ **Embedding-Level**	**42.0**	**55.8**	**27.4**	**36.6**	**42.2**	**54.4**	**38.2**	**52.1**
DPR	20.6	21.7	10.8	13.5	25.0	34.2	23.8	34.4
+ Prompt-Level	24.6	30.0	11.6	18.8	39.8	52.3	37.6	48.2
+ **Embedding-Level**	**29.8**	**36.3**	**16.8**	**21.0**	**43.0**	**54.4**	**38.0**	**51.9**
Contriever	22.6	35.4	16.6	20.7	25.8	32.8	25.2	34.2
+ Prompt-Level	30.2	47.3	17.6	25.9	39.8	52.2	34.6	48.7
+ **Embedding-Level**	**36.6**	**52.7**	**26.4**	**34.2**	**42.2**	**53.6**	**36.0**	**49.6**

(3) Effect of Exploration with Exploratory Embedding.



- Exploratory embedding increases the **activation rate** of neurons by roughly **2–4%**.

Dataset	Method	BM25	DPR	Contriever
		Candidates Coverage (%): ↑		
HotpotQA	EmbQA	**58.0**	**50.0**	**51.0**
	w/o exploratory	55.0	46.0	48.2
2Wiki	EmbQA	**51.0**	**43.6**	**31.6**
	w/o exploratory	47.0	40.2	28.2
NQ	EmbQA	**46.6**	**43.6**	**45.8**
	w/o exploratory	43.0	41.8	41.2
WebQ	EmbQA	**57.8**	**56.6**	**51.8**
	w/o exploratory	54.6	52.0	49.2