Zero-RAG: Towards Retrieval-Augmented Generation with Zero Redundant Knowledge

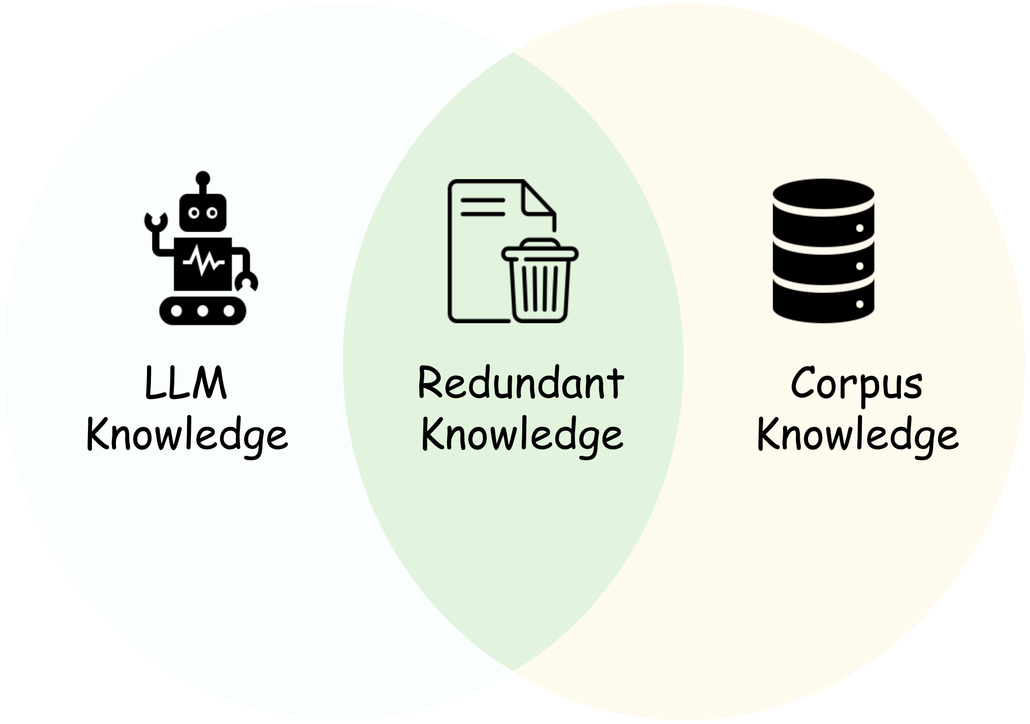
Qi LUO, Xiaonan LI, Junqi DAI, Shuang CHENG, Yining ZHENG, Xipeng QIU

Frontiers of Computer Science, DOI: 10.1007/s11704-026-50986-4

Problems & Ideas

- Problems of conventional stereo matching approaches:
  - Significant knowledge redundancy exists between external corpora and large language models, leading to unnecessary retrieval overhead.
  - Retrieving redundant knowledge is not only unnecessary but can degrade model performance

  Ideas: Zero-RAG identifies and removes knowledge that LLMs have already internalized, enabling faster and more efficient RAG systems without compromising accuracy.



Main Contributions

- Contributions:
  - First to explore RAG-oriented corpus pruning by removing knowledge redundancy between the corpus and the LLM, without compromising RAG's performance.
  - Proposed Mastery-Score metric to quantify how well an LLM has mastered each piece of knowledge.
  - Experimental results show that Zero-RAG prunes 30% of Wikipedia corpus, accelerates retrieval by 27%, while maintaining performance (average drop < 2 points).

Method	PopQA	HotpotQA	TriviaQA	EntityQuestions	AVG
		Llama3-70B			
Llama3-70B-Instruct	14.08	43.71	80.66	51.91	47.59
+ Retrieval	15.62	43.87	76.44	48.25	46.05
Noise-Tolerant Tuning	25.21	42.70	81.43	54.14	50.87
+ Retrieval	39.36	49.67	81.90	65.25	59.05
Zero-RAG (No Pruning)	31.72	45.00	81.80	60.82	54.84
Zero-RAG (- 10% Corpus)	30.81	43.84	81.50	58.92	53.77
Zero-RAG (- 30% Corpus)	30.67	43.30	81.00	57.82	53.20
Zero-RAG (- 50% Corpus)	30.32	41.93	80.53	56.11	52.22
Zero-RAG (- 70% Corpus)	29.48	40.69	80.40	55.41	51.50