RaLMSpec: Accelerating Retrieval-augmented Language Model Serving with Speculation

Carnegie Mellon University

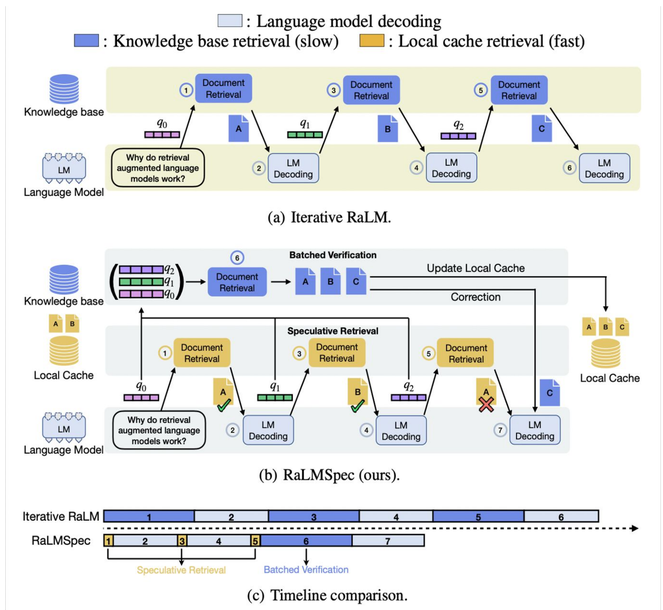
Lijie Yang *(School of Computer Science) Mentor: Prof. Zhihao Jia, Zhihao Zhang*

Introduction

- **Retrieval-augmented language models (RaLM) have demonstrated the potential to solve NLP tasks by incorporating a non-parametric knowledge base.**
- Existing RaLM methods can be categorized into two classes based on interaction with the knowledge base:
  (1) **One-shot**: retrieve **once** for each request
  (2) **Iterative**: **periodically** query the knowledge base
- Although iterative RaLM achieves better generative quality, frequent retrievals produces high retrieval overhead. This project RaLMSpec answers the following research question: **can we reduce the overhead of iterative RaLM without affecting generative quality?**

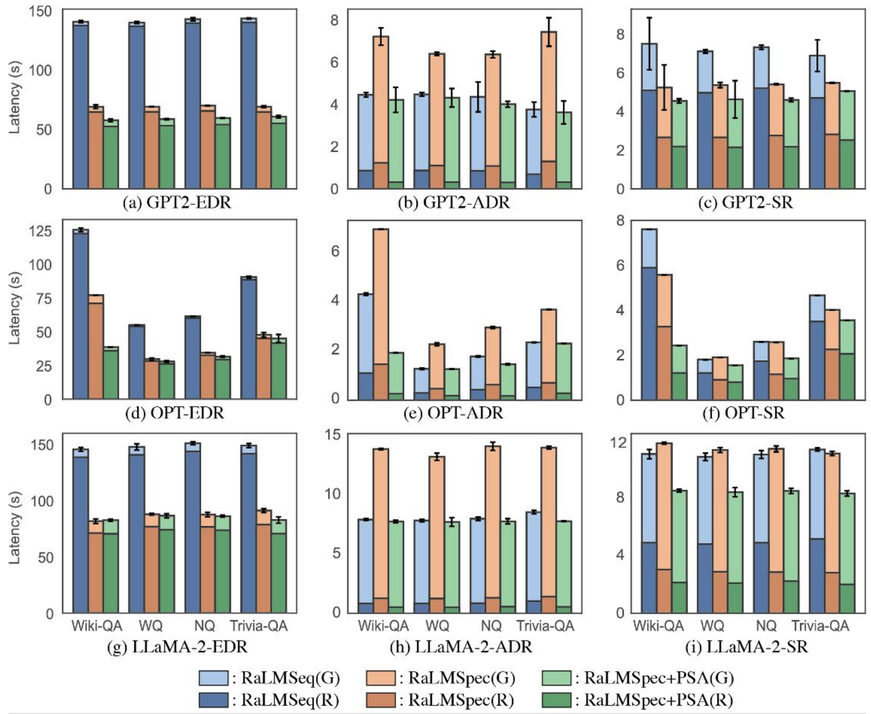
Methodology

- Leveraging the **temporal locality** (i.e., the same document can be retrieved multiple times of a generative task), RaLMSpec combines **cache-based speculative retrieval** with **batched verification** to **preserve the generation quality.**
- Incorporating **prefetching (P), optimal speculation stride scheduler (S),** and **asynchronous verification (A)** exploited the performance of RaLMSpec to the fullest.



Results

- **Latency summary** of RaLMSeq (Baseline), RaLMSpec, and RaLMSpec+PSA on GPT2-medium, OPT-1.3B, and LLaMA-2-7B over four QA datasets with exact dense (EDR), approximate dense (ADR), and sparse (SR) retrievers; G-generation, R-retrieval latency.



- **Ablation results of speed-up** (*) and (**) denote the most and the second most speed-up ratio

Retriever	Method	GPT2	OPT	LLaMA-2
EDR	RaLMSpec	2.04×	1.76×	1.70×
	RaLMSpec+P	2.10×	2.16×(**)	1.75×(**)
	RaLMSpec+S	2.26×(**)	2.15×	1.69×
	RaLMSpec+A	2.03×	1.74×	1.74×
	RaLMSpec+PSA	2.39×(*)	2.32×(*)	1.75×(*)
ADR	RaLMSpec	0.62×	0.61×	0.58×
	RaLMSpec+P	0.59×	0.76×	0.58×
	RaLMSpec+S	0.92×(**)	1.17×(**)	1.01×(**)
	RaLMSpec+A	0.66×	0.46×	0.55×
	RaLMSpec+PSA	1.05×(*)	1.39×(*)	1.04×(*)
SR	RaLMSpec	1.34×	1.18×	0.97×
	RaLMSpec+P	1.39×	1.42×	0.98×
	RaLMSpec+S	1.32×	1.52×(**)	1.05×(**)
	RaLMSpec+A	1.41×(**)	1.27×	1.01×
	RaLMSpec+PSA	1.53×(*)	1.77×(*)	1.31×(*)

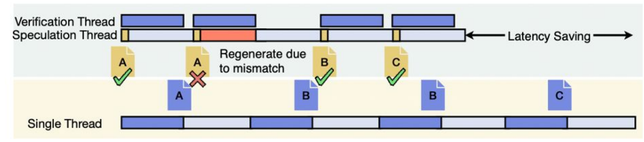
Discussion

The following are key features behind the speedup of RaLMSpec as a speculation-inspired framework that accelerates the serving of generic retrieval augmented generation.

- **Stride: the number of speculation steps performed before a verification step**,

$$\hat{\gamma}(X) = \frac{\sum_t M(s(t), X)}{\sum_t M(s(t), X) + \sum_t \mathbb{1}(M(s(t), X) < s(t))}$$

- **Asynchronous verification: launch a new speculation step asynchronously** while the verification of the previous step occurs.



Takeaway

- RaLMSpec effectively **reduces the retrieval overhead** of iterative RaLM with **batch verification** and **cache-based speculation while maintaining the same generation quality**.
- Extensive evaluations demonstrate that RaLMSpec can achieve a speed-up ratio of **1.75-2.39X (EDR)**, **1.04-1.39X (ADR)**, and **1.31-1.77X (SR)**

Future Work

- Run additional experiments on larger main-stream models (such as LLaMA-2-70B)
- Investigate the workload in approximate dense retriever and sparse retriever

Contact
lijiey@andrew.cmu.edu