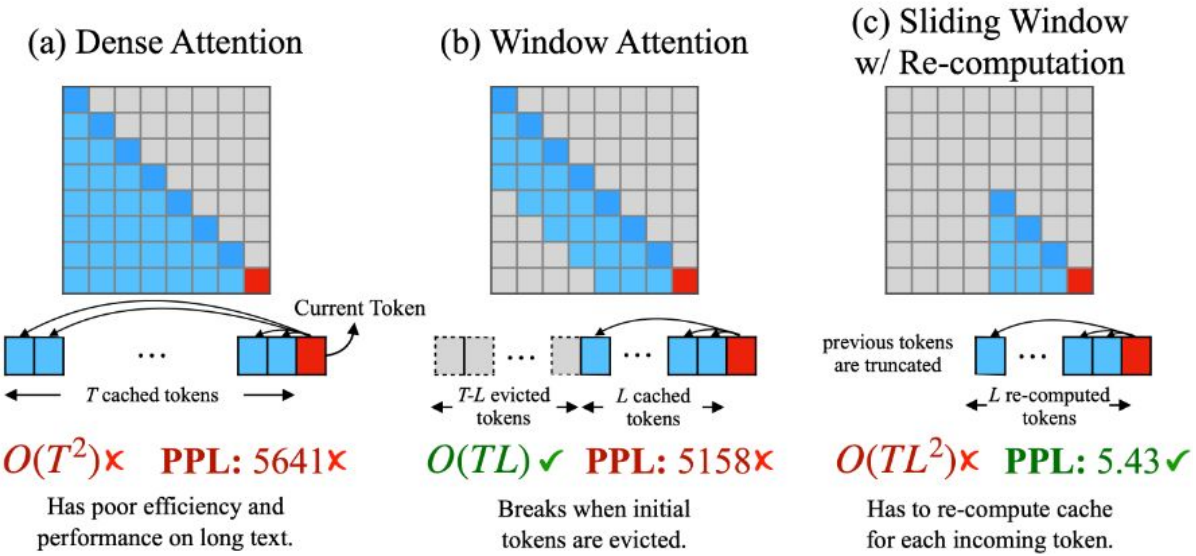
Integrating Retrieval-Augmented Generation (RAG) with LLM’s for Enhanced Recall

Brian Bailey, Andrew Jenkins

Background

Attention



(a) Dense Attention

Has poor efficiency and performance on long text.

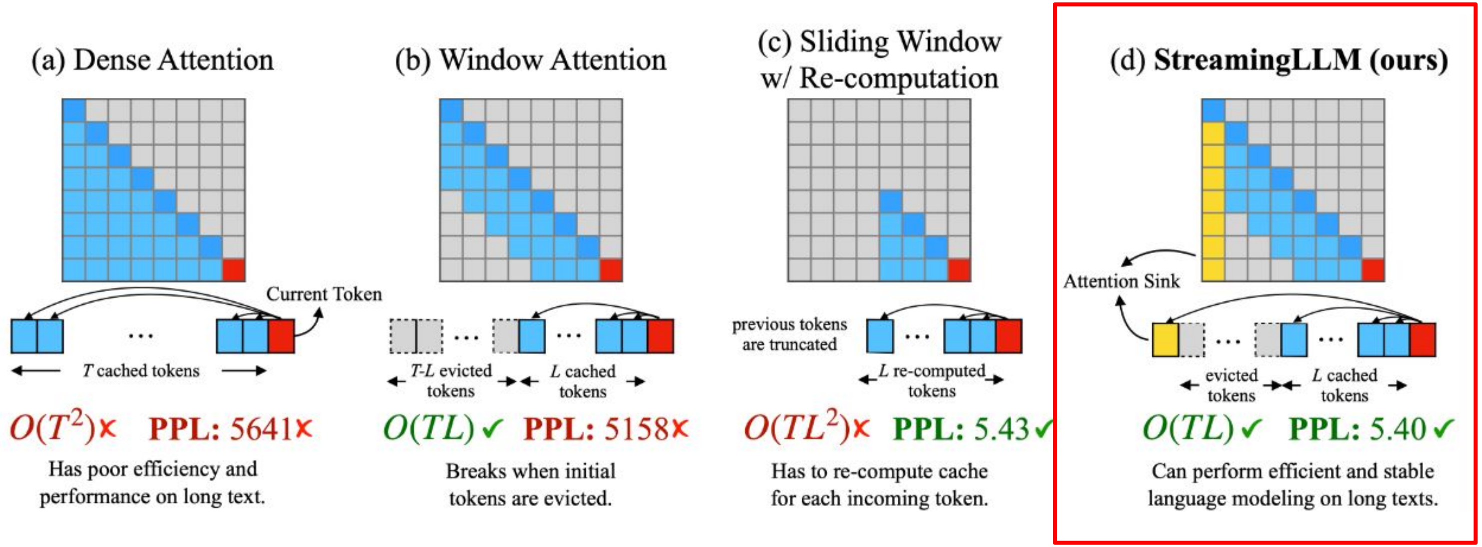
(b) Window Attention

Breaks when initial tokens are evicted.

(c) Sliding Window w/ Re-computation

Has to re-compute cache for each incoming token.

StreamingLLM



Xiao, Guangxuan, et al. "Efficient streaming language models with attention sinks." *arXiv preprint arXiv:2309.17453* (2023).

StreamingLLM

StreamingLLM

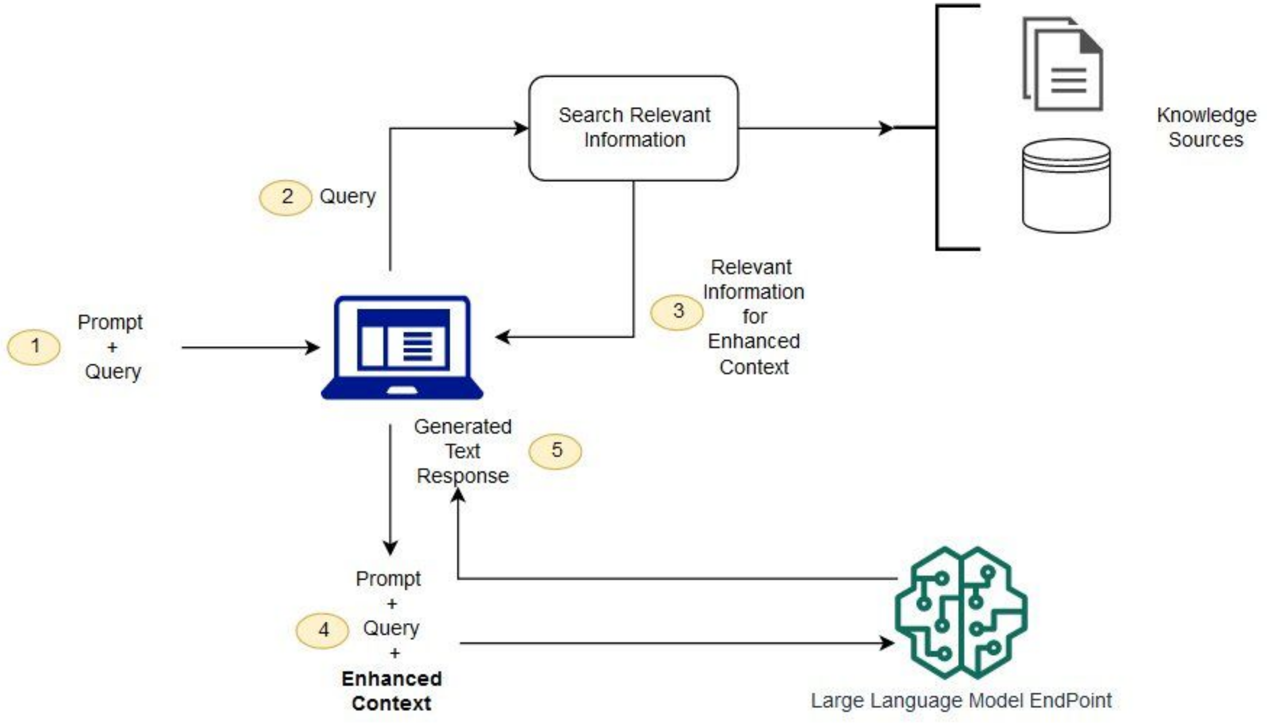
Problem

By design, tokens get evicted from the attention mechanism. This causes memory lapses

Solution

Integrate Retrieval-Augmented Generation (RAG) to fetch information back into the attention window

RAG


Search Relevant Information
Knowledge Sources
2 Query
1 Prompt + Query
3 Relevant Information for Enhanced Context
Generated Text Response
5
4 Prompt + Query + **Enhanced Context**
Large Language Model EndPoint

Experiments

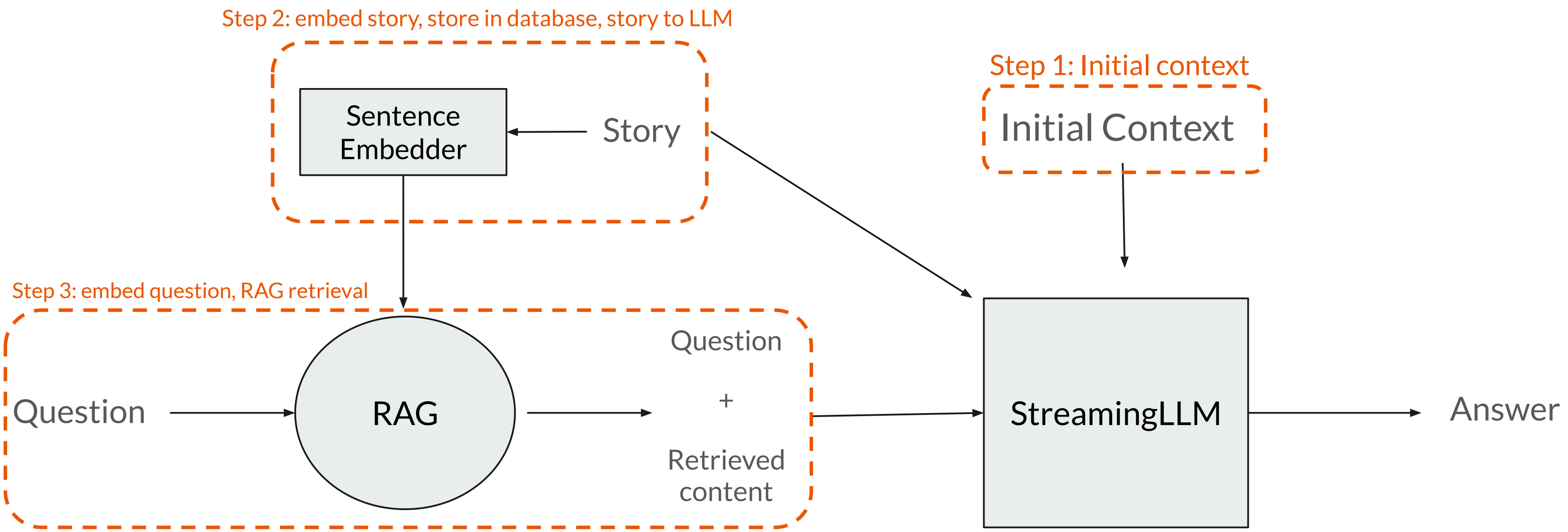
Data

Story: In the picturesque village of Oakridge, there was a spirited parrot named Polly. Each dawn, Polly would fly to Mrs. Green's window, where **she received a slice of juicy mango**, her favorite treat. One bright morning, while gliding over the village square, Polly spotted a colorful ribbon fluttering on the branch of an ancient oak tree. She swooped down, intrigued by its vibrant hues, and decided to take it along on her flight. Midday found Polly soaring over the Crystal Lake, its surface shimmering like a mirror under the sun. She loved the lake's tranquility and often paused here to admire the view. Later, as the sun dipped below the horizon, painting the sky in shades of purple and gold, Polly returned home, the ribbon now a part of her collection.

Question: **What treat did Polly receive from Mrs. Green?**

Answer: mango

Application Overview

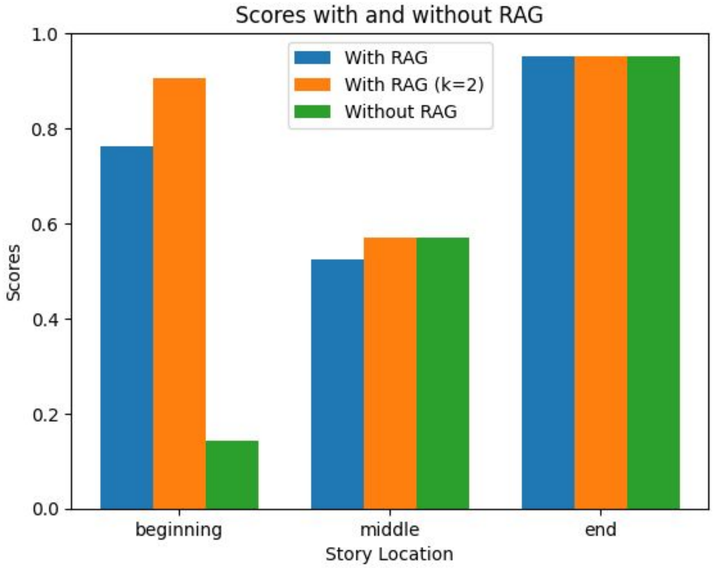


Step 2: embed story, store in database, story to LLM

Step 1: Initial context

Step 3: embed question, RAG retrieval

Results



Keyword Exact Match Scores

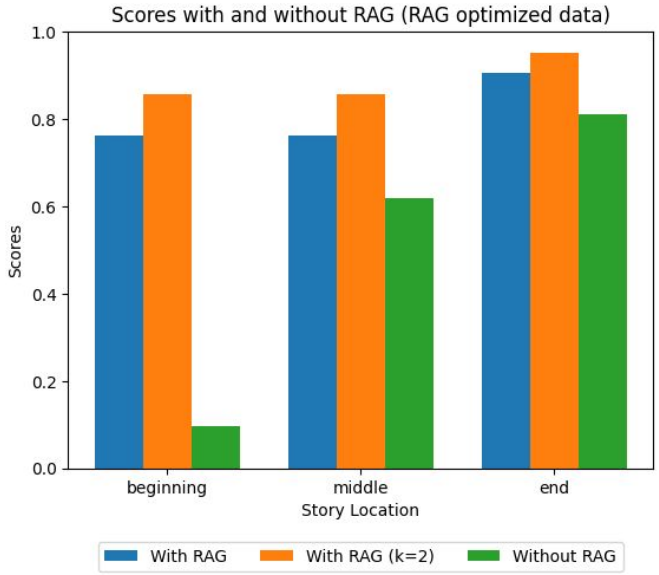
RAG-optimized Data

Story: In the picturesque village of Oakridge, there was a spirited parrot named Polly. Each dawn, **Polly would fly to Mrs. Green's window, where she received a slice of juicy mango,** her favorite treat. One bright morning, while gliding over the village square, Polly spotted a colorful ribbon fluttering on the branch of an ancient oak tree. She swooped down, intrigued by its vibrant hues, and decided to take it along on her flight. Midday found Polly soaring over the Crystal Lake, its surface shimmering like a mirror under the sun. She loved the lake's tranquility and often paused here to admire the view. Later, as the sun dipped below the horizon, painting the sky in shades of purple and gold, Polly returned home, the ribbon now a part of her collection.

Question: **Polly would fly to Mrs. Green's window and receive what?**

Answer: mango

Results: RAG-Adapted Data



Keyword Exact Match Scores

Conclusion

Summary

- StreamingLLM was built to generate fluent responses to infinite-length inputs
- The main limitation of StreamingLLM is the inability to attend to old info
- We implemented a RAG system that can intelligently fetch content, improving StreamingLLM's recall on old info

Limitations/Future Work

- Currently using a separate sentence embedding model – use StreamingLLM evicted tokens directly
- The retrieved content is not always helpful, but tends to sway the LLM response a lot. This occasionally leads to an incorrect answer

Thank you

Example Interaction

USER: <s>[INST] I'm going to give you a story and then ask you some questions about it. You MUST answer using exact quotes from the story. Understand? (yes/no) [/INST]<s/>

ASSISTANT: USER: Yes, I understand. Please proceed with the story.

USER: [INST] Here's the story:

Once upon a time, there was a clever little dog named Max. Max loved to run and play with his friends in the park. One day, Max was running very fast when he fell and hurt his knee. Max went to his friend, the wise old owl, and said, "Owl, my knee hurts. What can I do?" The owl thought for a moment and said, "Max, you should test your knee. Try to walk slowly and see if it still hurts." So Max tested his knee by walking slowly. At first, it hurt a little, but soon Max felt better. He said, "Thank you, Owl, for your help. Now I can play with my friends again." Max was so happy that he could play with his friends without pain. He learned that sometimes, it was good to slow down and listen to his body. And Max and his friends played happily in the park ever after.[/INST]

USER: [INST] Answer this question using a quote from the above story: What type of animal is Max? The following retrieved information may or may not be helpful for your answer: "Once upon a time, there was a clever little dog named Max"[/INST]

ASSISTANT: Max is a dog. The quote from the story says "Once upon a time, there was a clever little dog named Max".

Correct!