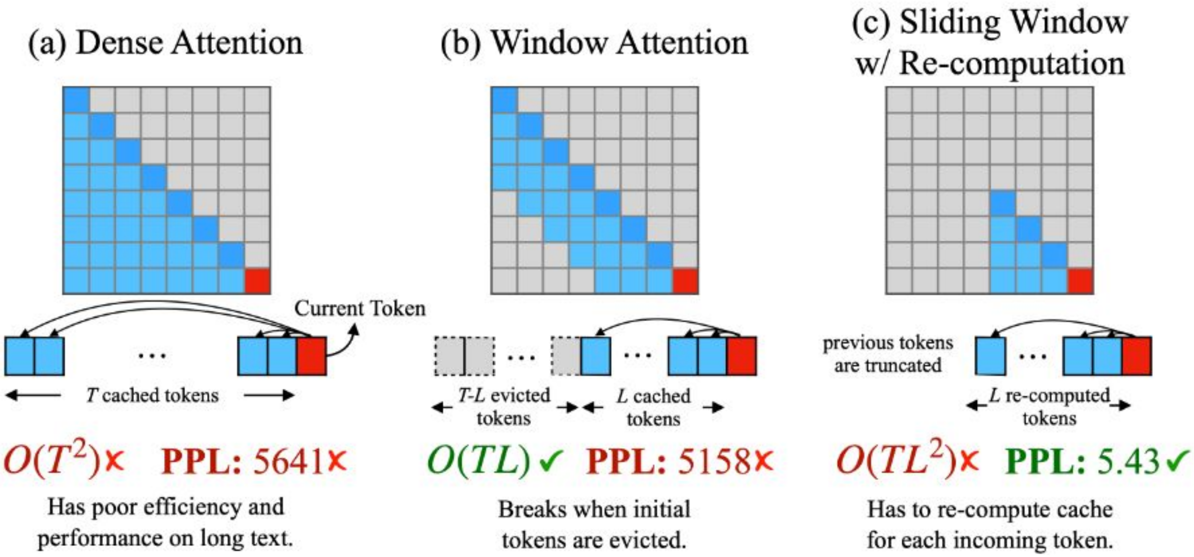
Integrating Retrieval-Augmented Generation (RAG) with LLM's for Enhanced Recall

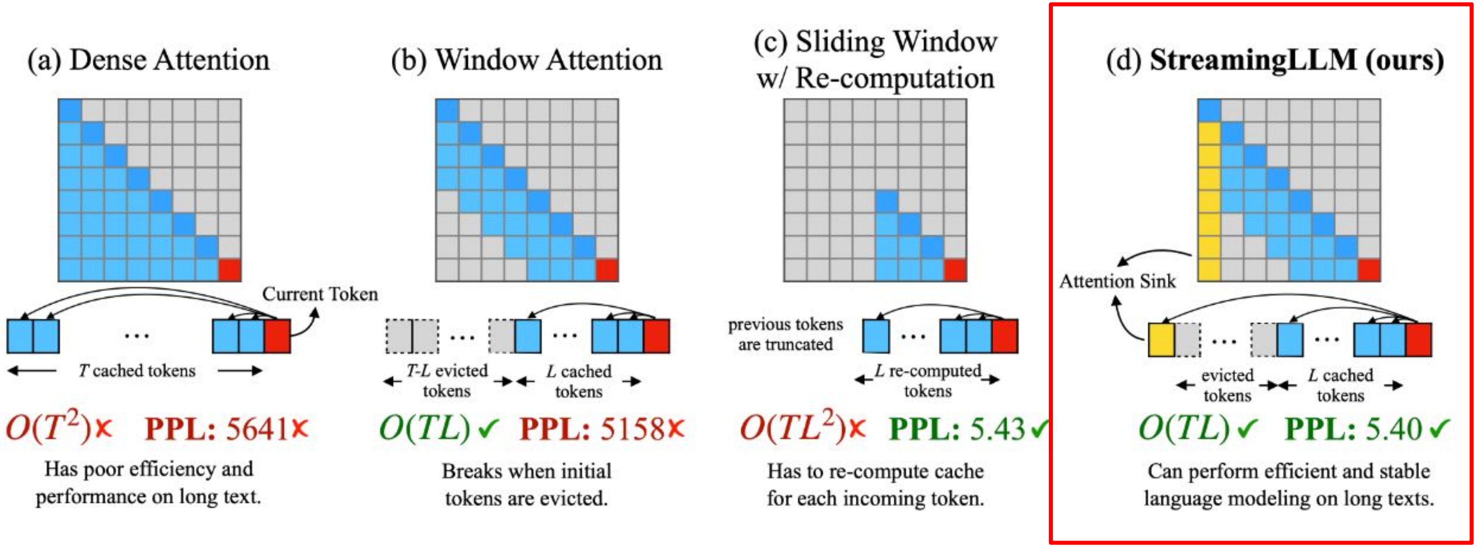
Brian Bailey, Andrew Jenkins

Background

Attention


(a) Dense Attention
Current Token
...
T cached tokens
$O(T^2)$✗ PPL: 5641✗
Has poor efficiency and performance on long text.
(b) Window Attention
...
...
T-L evicted tokens
L cached tokens
$O(TL)$✓ PPL: 5158✗
Breaks when initial tokens are evicted.
(c) Sliding Window w/ Re-computation
previous tokens are truncated
...
L re-computed tokens
$O(TL^2)$✗ PPL: 5.43✓
Has to re-compute cache for each incoming token.

StreamingLLM



StreamingLLM

StreamingLLM

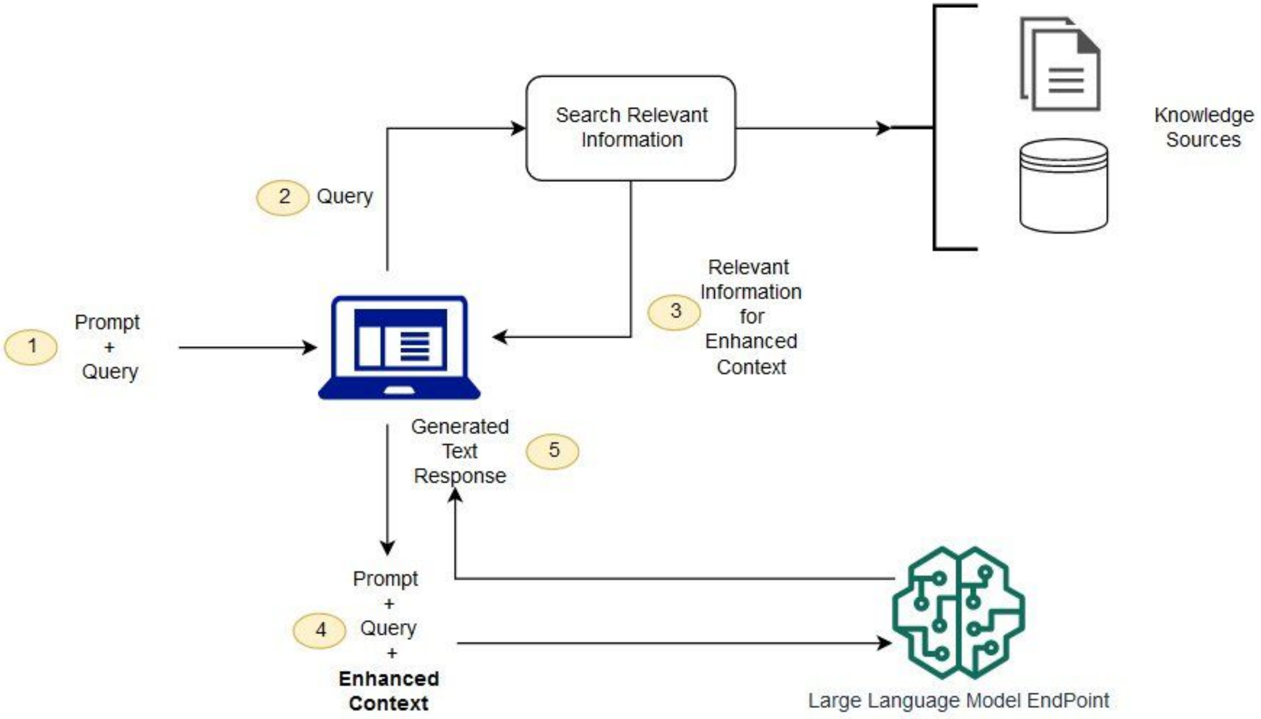
Problem

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

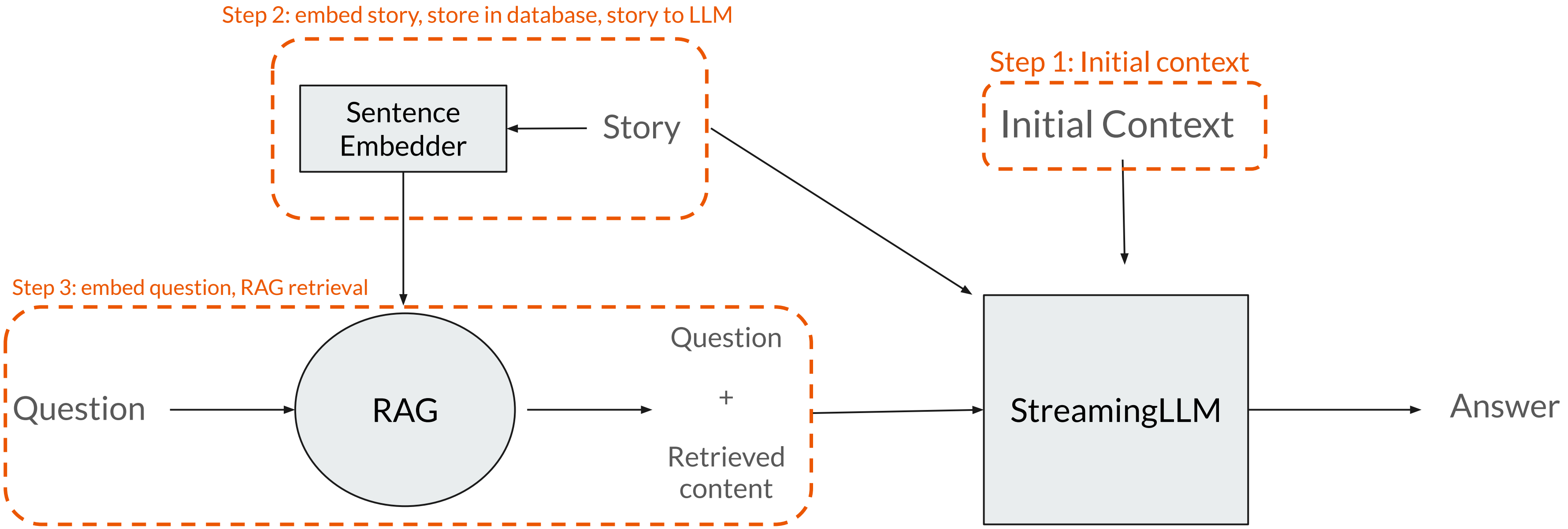
Data

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.

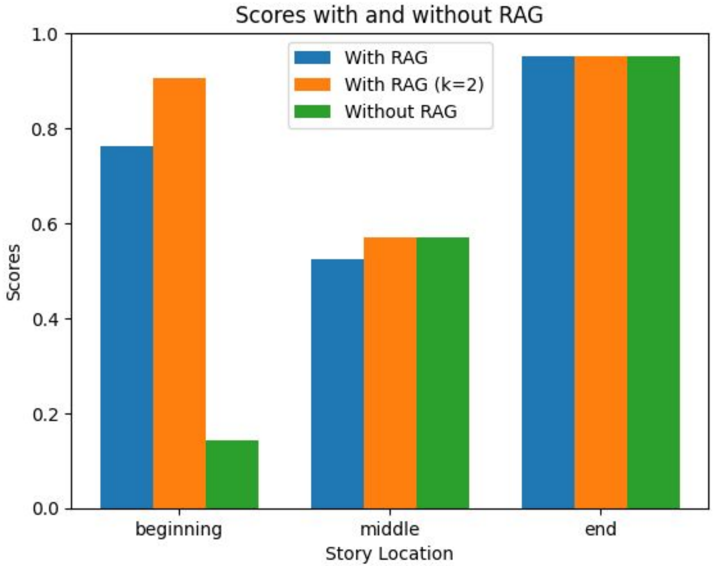
Question: **What treat did Polly receive from Mrs. Green?**

Answer: mango

Application Overview



Results



Keyword Exact Match Scores

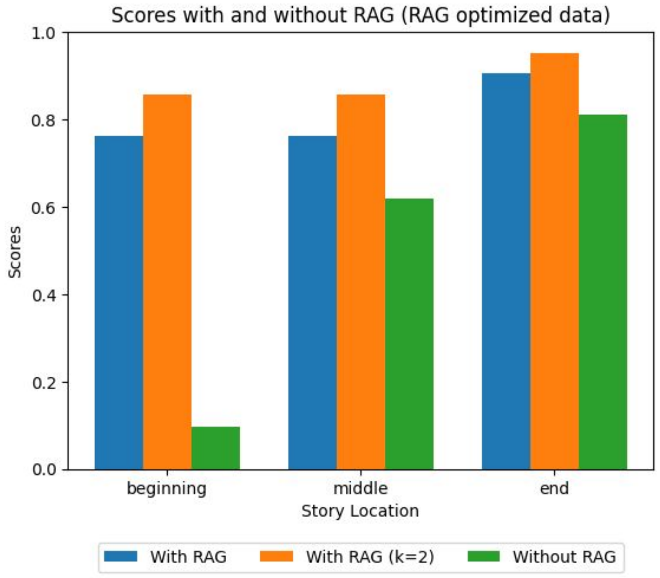
RAG-optimized Data

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

RAG

Story → [Sentence 1,
Sentence 2,
...]

Transformers

Example Test

example of inputs, model outputs, rag return, answer