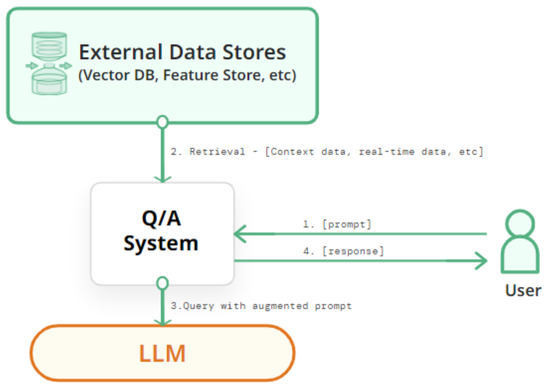
Enhancing Long Context Performance in LLMs Through Inner Loop Query Mechanism

Yimin Tang**, Yurong Xu*, Ning* Yan*, Masood Mortazavi*

*Futurewei Technologies **University of Southern California

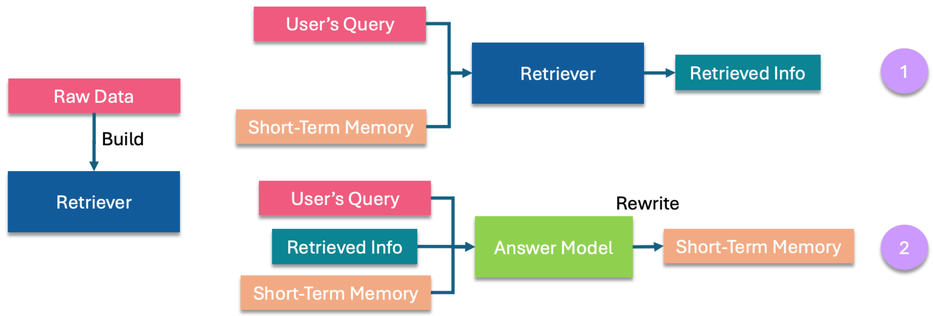
1. Background

RAG-based retrieval based on the initial query may not work well when dealing with questions posted regarding large text. Such cases require deeper reasoning.



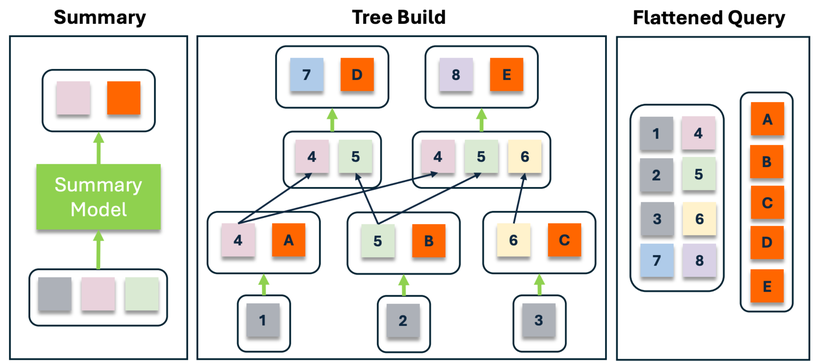
2. Inner-Loop Memory-Augmented Tree Retrieval (ILM-TR)

Using information retrieved from the RAG system, our model drives an LLM to generate and store text in an area named Short-Term Memory (STM). This retrieval process repeats until the text in STM converges.



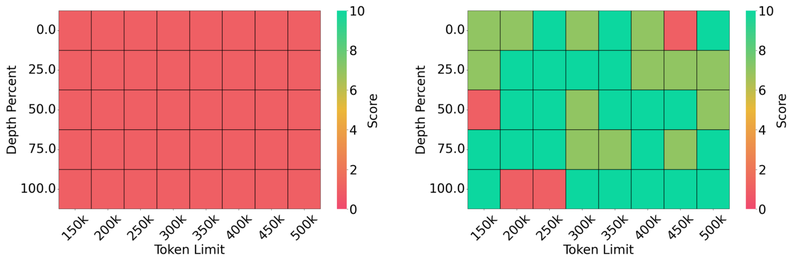
3. Retriever Structure

Our basic structure mimics RAPTOR [3]: original text (**gray**), surprising information (orange), summary information (other colors). In the query process, all blocks in the tree will be stored in a lookup table, and the best fit will be returned based on vector distance from the query text.

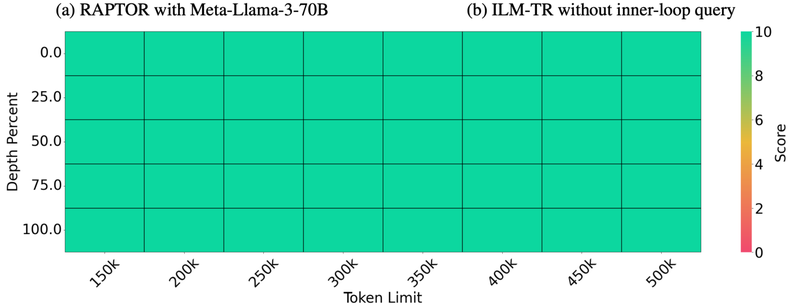


4. M-NIAH Experiment

Several sentences are inserted into a specific area of a given long context. The question is related to all the inserted sentences, and the model is expected to retrieve all necessary information across these sentences.



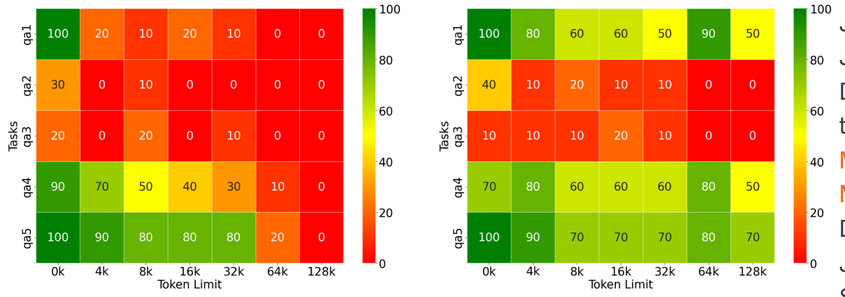
(a) RAPTOR with Meta-Llama-3-70B (b) ILM-TR without inner-loop query



(c) ILM-TR

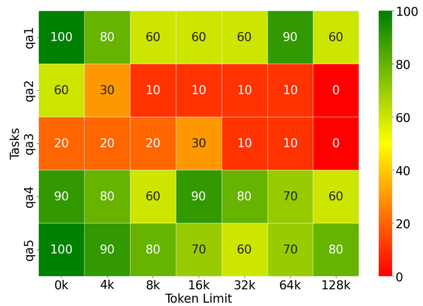
5. BABILong Experiment

The BABILong [2] test is similar to the M-NIAH test but involves sentences with more complex logical relationships.



(a) RAPTOR with Meta-Llama-3-70B

(b) ILM-TR without inner-loop query



(c) ILM-TR

6. Examples

[BABILong Needles]:
Daniel grabbed the milk.
Mary picked up the apple.
John moved to the bedroom.
John went to the bathroom.
Daniel discarded the milk there.
Mary moved to the kitchen.
Mary journeyed to the office.
Daniel got the milk.
John moved to the garden.
Sandra travelled to the kitchen.
Mary put down the apple.
John took the football.
[Question]: Where was the apple before the office?
[True Answer]: kitchen

[Round 1]:
The apple was at office. We need to find where the apple was before the office.
[Round 2]:
The apple was at office. Mary put down the apple at office. We need to determine where Mary was before she placed the apple down.
[Round 3]:
The apple was at office. Mary put down the apple at office, but before that, she was in the kitchen.
[Round 4]:
Mary put down the apple at office, but before that, she was in the kitchen. The best answer to the question ""Where was the apple before the office?"" is:
The kitchen.
[Round 5]:
Based on the given context, the best answer to the question ""Where was the apple before the office?"" is:
The kitchen.

References

[1] Tang, Yimin, et al. "Enhancing Long Context Performance in LLMs Through Inner Loop Query Mechanism." *arXiv preprint arXiv:2410.12859* (2024).

[2] Kuratov, Yuri, et al. "BABILong: Testing the Limits of LLMs with Long Context Reasoning-in-a-Haystack." *arXiv preprint arXiv:2406.10149* (2024).

[3] Sarthi, Parth, et al. "Raptor: Recursive abstractive processing for tree-organized retrieval." *arXiv preprint arXiv:2401.18059* (2024).