PECAN: LLM-Guided Dynamic Progress Control with Attention-Guided Hierarchical Weighted Graph for Long-Document QA

Xinyu Wang, Yanzheng Xiang, Lin Gui, and Yulan He

Scan to check the paper!

Our method

- In this paper, we combine the high accuracy of LLMs with the efficiency of RAG and propose LLM-Guided Dynamic **P**rogr**e**ss **C**ontrol with **A**ttentio**n**-Based Hierarchical Weighted Graph (**PECAN**).
- (1) LLM-Guided Dynamic Progress Control: We leverage LLMs to dynamically control the retrieval process, adjusting the amount of retrieved information based on different queries to achieve a better balance of effectiveness and efficiency.
- (2) Attention-Guided Retrieval: We propose a novel retrieval method that constructs a hierarchical graph where edges are derived by LLM attention weights.

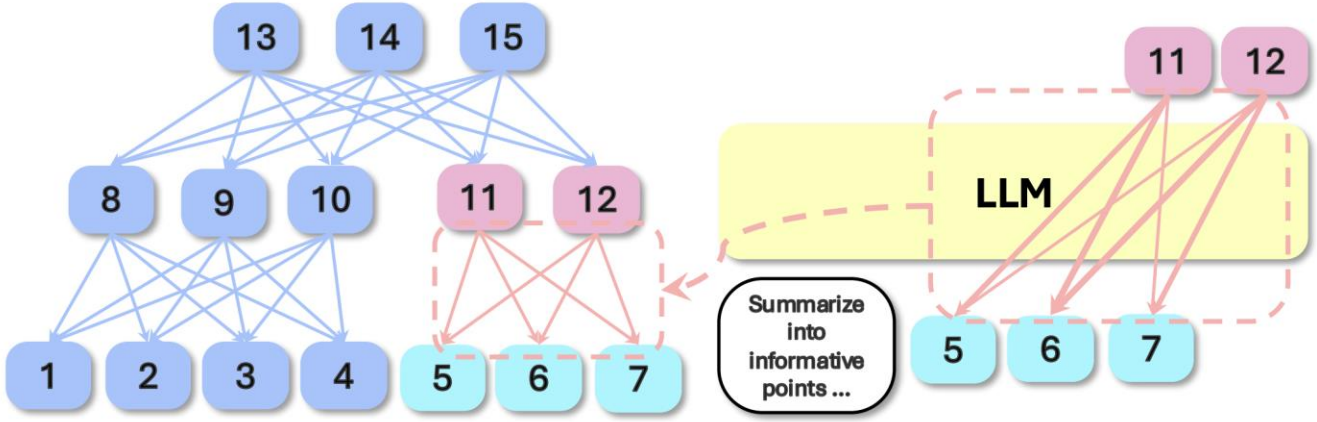
Overview

PECAN consists of two main steps:

- Attention Graph Construction: we utilize the LLM's attention weights to build a Hierarchical Weighted Directed Acyclic Graph (HWDAG) from documents. This is a one-time preprocessing step for each document, after which it can be reused for any query to that document
- Dynamic Graph Search: we dynamically control the volume of retrieved nodes and perform a search guided by the LLM.

Attention Graph Construction

- We iteratively summarize nodes to obtain the higher-level nodes.
- Each node typically summarizing one or a few events.
- The edge weights are derived from the LLM's attention during summarization.



11 SpongeBob SquarePants is an American animated television series created by Stephen Hillenburg for Nickelodeon.

12 "SpongeBob SquarePants" is set in the fictional underwater city of Bikini Bottom and centers on the adventures of SpongeBob SquarePants, an anthropomorphic sea sponge.

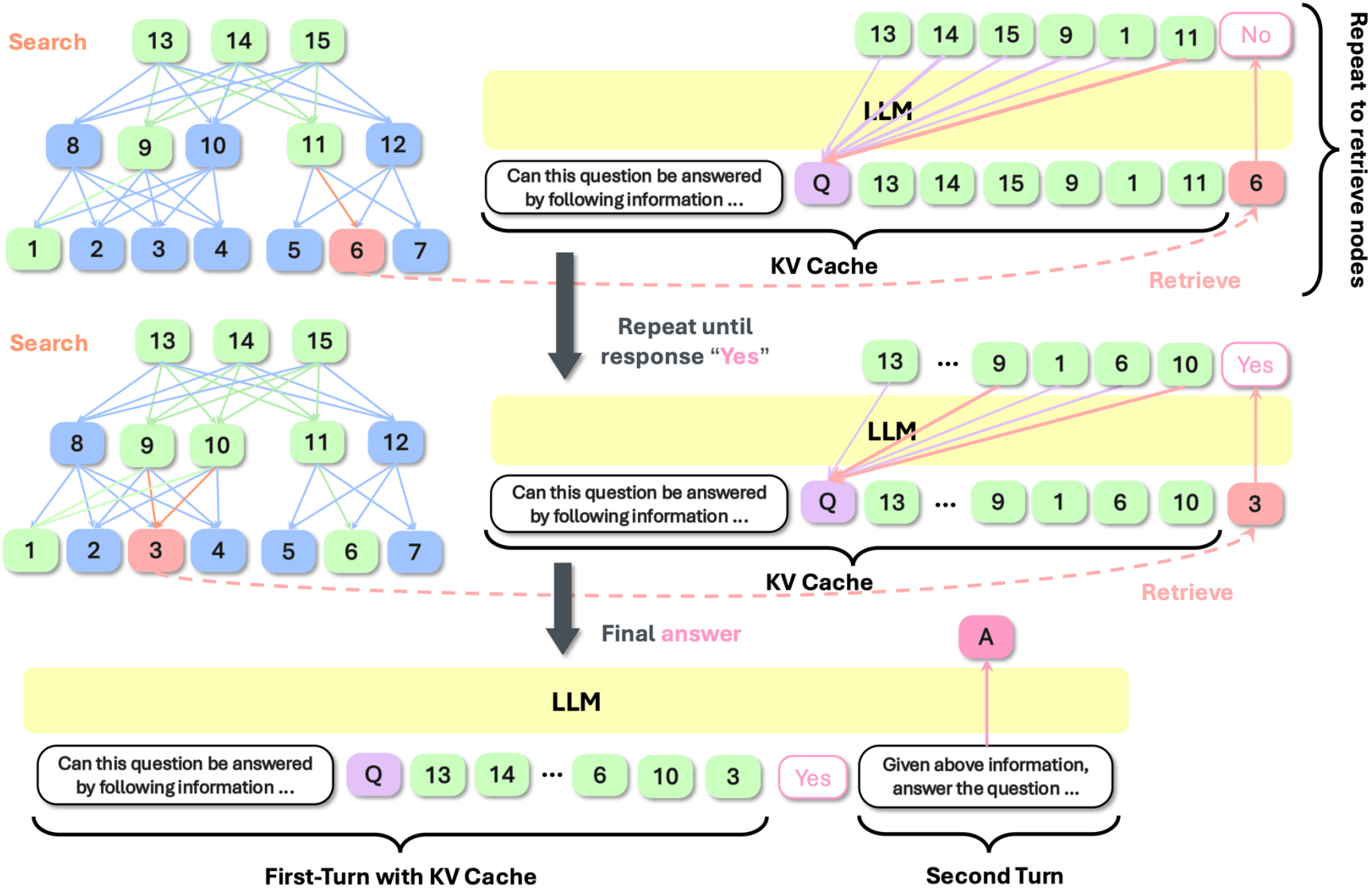
5 American animated television series "SpongeBob SquarePants" ... an anthropomorphic sea sponge named SpongeBob SquarePants, attempting to get a job at a local restaurant called the Krusty Krab. However, he is tasked to find a seemingly non-existent ...

6 "SpongeBob SquarePants" is an American animated television series created by marine biologist and animator Stephen Hillenburg ... centers on the adventures of SpongeBob SquarePants, an over-optimistic sea sponge that annoys other characters...

7 "SpongeBob SquarePants" chronicles the adventures and endeavors of the title character and his various friends in the fictional underwater city of Bikini Bottom ... show originated in an unpublished, educational comic book ...

Dynamic Graph Search

- If a node containing an event is strongly correlated with the query, then the details about that event are likely to be more useful in answering the query.
- At each iteration, a node is retrieved based on attention weights.
- LLM determines whether sufficient nodes have been gathered to answer the query.
- This procedure dynamically adapts to different query.



Experiments

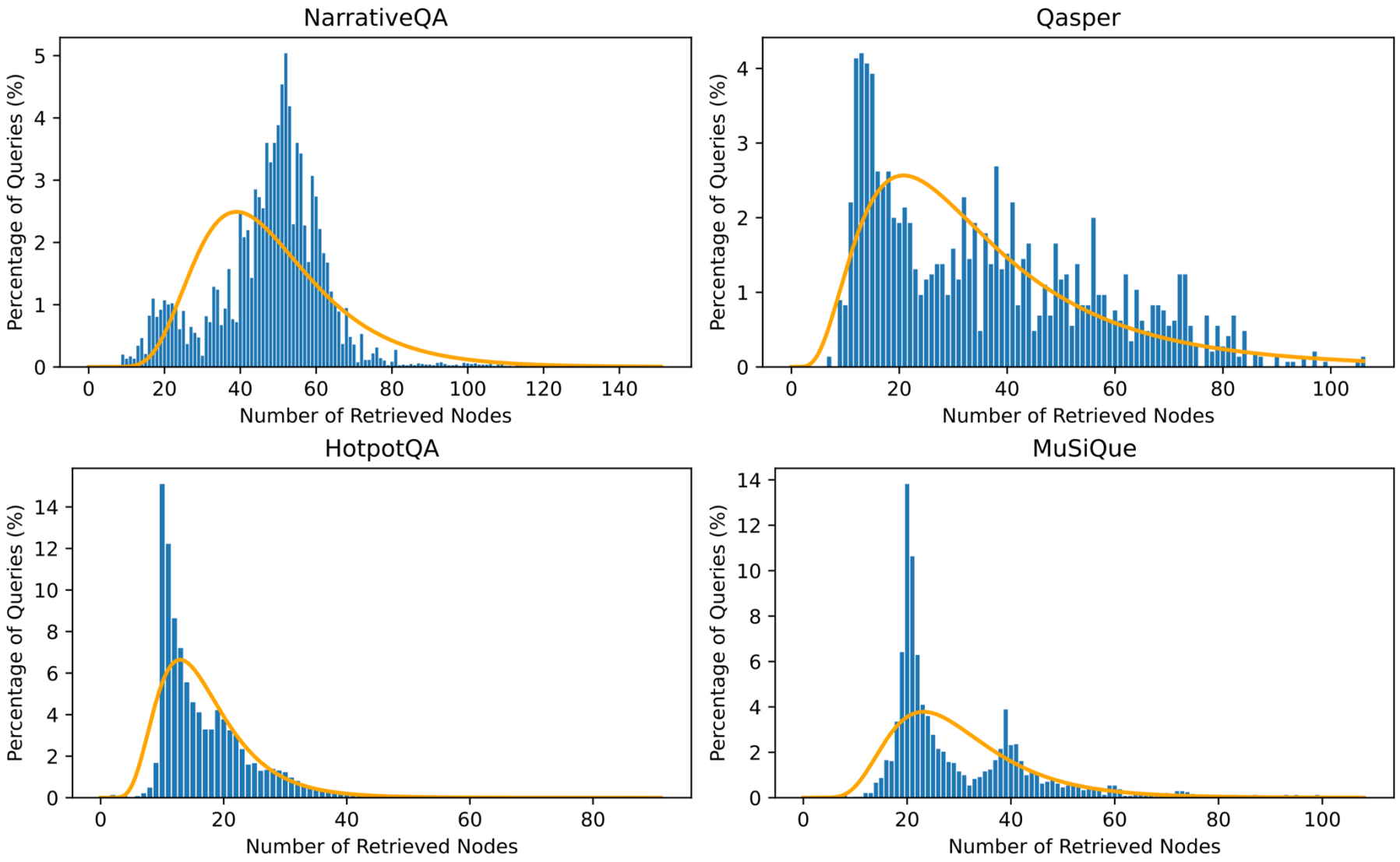
- We conduct experiments of baselines and PECAN on NarrativeQA, Qasper, HotpotQA, and MuSiQue.
- We include a Top-X setting to match the TFLOPs of PECAN for a fair comparison.

Method	NarrativeQA				Qasper			
	F1	ROUGE-L	TFLOPs	Ratio	F1	ROUGE-L	TFLOPs	Ratio
BM25 Top-5	52.7	51.8	26.7	0.86x	41.0	39.6	26.3	0.39x
SBERT Top-5	36.5	35.8	26.8	0.86x	44.4	42.4	26.0	0.39x
Dragon Top-5	53.8	52.9	26.9	0.87x	43.0	41.4	24.5	0.36x
MeMWalker	11.2	9.8	353.8	11.41x	39.0	36.8	123.9	1.85x
RAPTOR-TT	40.6	39.8	20.3	0.65x	42.1	40.1	17.7	0.26x
RAPTOR-CT Top-5	48.6	47.8	17.9	0.58x	44.6	42.7	16.6	0.25x
LongLLMLingua	50.5	49.5	1789.4	57.72x	43.2	43.0	159.7	2.39x
BM25 Top-X	53.7	52.9	37.5	1.21x	47.0	45.1	69.3	1.04x
SBERT Top-X	39.5	38.8	37.5	1.21x	46.6	44.5	68.9	1.03x
Dragon Top-X	55.1	54.2	37.5	1.21x	46.9	44.8	67.0	1.00x
RAPTOR-CT Top-X	52.0	51.2	35.1	1.13x	46.9	44.7	67.3	1.01x
Llama-3.1-8B	53.7	52.6	3361.9	108.45x	49.4	47.6	92.5	1.38x
PECAN	**61.1**	**60.2**	31.0	1.00x	**49.7**	**47.9**	66.9	1.00x

Method	HotpotQA				MuSiQue			
	F1	ROUGE-L	TFLOPs	Ratio	F1	ROUGE-L	TFLOPs	Ratio
BM25 Top-5	40.8	40.9	22.9	1.43x	28.7	28.7	26.3	0.85x
SBERT Top-5	40.9	40.8	22.6	1.41x	30.7	30.8	26.1	0.84x
Dragon Top-5	39.7	39.6	23.3	1.46x	28.5	28.4	28.1	0.91x
MeMWalker	39.7	38.9	93.4	5.84x	24.0	23.5	175.7	5.69x
RAPTOR-TT	38.6	38.5	8.4	0.53x	29.3	29.3	12.6	0.41x
RAPTOR-CT Top-5	40.9	40.4	15.3	0.96x	31.5	31.5	16.1	0.52x
LongLLMLingua	43.4	**43.5**	43.6	2.73x	34.5	34.4	78.9	2.55x
BM25 Top-X	40.7	40.8	20.0	1.25x	31.8	31.7	35.6	1.15x
SBERT Top-X	40.8	40.7	19.6	1.23x	32.5	32.5	35.6	1.15x
Dragon Top-X	39.2	39.1	20.6	1.29x	30.2	30.1	38.0	1.23x
RAPTOR-CT Top-X	40.7	40.7	17.9	1.12x	35.4	35.2	32.2	1.04x
Llama-3.1-8B	41.3	41.2	23.7	1.48x	35.8	35.7	40.6	1.31x
PECAN	**43.5**	**43.5**	16.0	1.00x	**36.9**	**36.8**	30.9	1.00x

Dynamic Retrieval Analysis

- Frequency distribution of the number of nodes retrieved per query.
- The x-axis represents the number of nodes retrieved per query.
- The y-axis indicates the percentage of queries retrieving that number of nodes.



Inference with Smaller Models

- We further investigate whether the effectiveness-efficiency tradeoff can be improved when the model used for graph search is smaller than the one used for graph construction.
- We experiment with combinations of of 8B, 1B, and 3B.

Graph Construction	Graph Search	NarrativeQA	Qasper	HotpotQA	MuSiQue
Llama-3.1-8B	Llama-3.1-8B	61.1	49.7	43.5	36.9
Llama-3.2-3B	Llama-3.2-3B	56.7	47.6	40.0	29.8
Llama-3.1-8B	Llama-3.2-3B	60.6	49.6	40.8	31.3
Llama-3.2-1B	Llama-3.2-1B	28.8	33.5	27.5	15.3
Llama-3.1-8B	Llama-3.2-1B	41.8	37.2	28.5	15.6

Thank you!

Scan to check the paper!