Recursive Abstractive Processing for Tree-Organized Retrieval (RAPTOR):

One of the limitations of RAG methods is that they can only retrieve short contiguous chunks from a retrieval corpus, which restricts their ability to provide a good understanding of the overall document context.

Recursive Abstractive Processing for Tree-Organized Retrieval (RAPTOR), is a novel approach that recursively embeds, clusters, and summarizes text chunks. RAPTOR constructs a tree with varying summarization levels, building up from the bottom.

During inference, the RAPTOR model retrieves information from this tree, enabling integration across lengthy documents at different levels of abstraction.

Experimental results demonstrate that retrieval with recursive summaries yields significant improvements over traditional retrieval-augmented language models across multiple tasks.

Link to RAPTOR paper: https://lnkd.in/dYWVDpnR

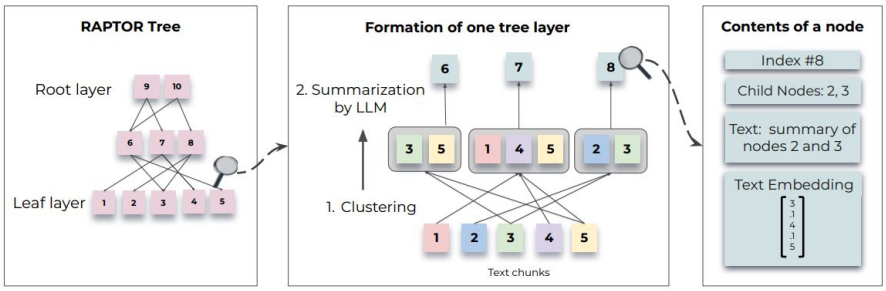

Figure 1: **Tree construction process:** RAPTOR recursively clusters chunks of text based on their vector embeddings and generates text summaries of those clusters, constructing a tree from the bottom up. Nodes clustered together are siblings; a parent node contains the text summary of that cluster.